Nommage de sens à l'aide des vecteurs conceptuels

Word Sense Naming with Conceptual Vectors

Fabien JALABERT[1&2] *
fabien.jalabert@ema.fr

Mathieu Lafourcade[1]
lafourca@lirmm.fr

[1] LIRMM
161 rue Ada
34 392 - Montpellier Cedex 5
www.lirmm.fr

[2] LGI2P - EMA
Parc Scientifique Georges Besse
30 035 Nîmes Cedex 1
www.lgi2p.ema.fr

Résumé

Dans le cadre de la recherche en sémantique lexicale, nous utilisons le modèle des vecteurs conceptuels pour représenter les aspects thématiques des mots. La base vectorielle est construite à partir de définitions provenant de différentes sources lexicales, ce qui permet statistiquement de tempérer les diverses incohérences locales. Pour désigner le sens obtenu après un regroupement des définitions, nous utilisons un identifiant qui entraîne certaines contraintes. En particulier, un "cluster" de définitions est désigné par une liste de références vers différentes définitions de la multi-source. D'autre part, le contrôle de la qualité d'une classification ou plus généralement d'une désambiguïsation de sens impose de faire référence en permanence au lexique source. Nous proposons donc de nommer un sens à l'aide d'un autre terme du lexique. L'annotation est un outil léger et efficace qui est essentiellement une association d'idées que l'on peut extraire de toute base de connaissance linguistique. Les annotations obtenues peuvent finalement constituer une nouvelle source d'apprentissage pour la base de vecteurs conceptuels.

Mots Clef

Traitement automatique du langage naturel, désambiguïsation sémantique lexicale, annotation sémantique lexicale, modèle des vecteurs conceptuels.

Abstract

In the research framework in meaning representation in NLP, we focus our attention on thematic aspects and conceptual vectors. This vectorial base is built upon a morphosyntactic analysis of several lexical resources to reduce isolated problems. Also a meaning is a cluster of definitions that are pointed by an Id number. To check the results of an automatic clustering or a word sense disambiguation, we must continuously refer to the source dictionary. In this article, we describe a method for naming a word sense by a term of the vocabulary. This kind of annotation is a light and efficient method that uses meanings associations someone or something can extract from any lexical knowledge base. Finally, the annotations should become a new lexical learning resource to improve the vectorial base.

Keywords

Natural Language Processing, Word Sense Disambiguation, Word Sense Tagging, Conceptual Vector Model.

Introduction

Dans le cadre de la recherche en sémantique lexicale, l'équipe TAL du LIRMM développe actuellement un système d'analyse des aspects thématiques des textes et de désambiguïsation lexicale basé sur les vecteurs conceptuels. Les vecteurs représentent les idées associées à tout segment textuel (mots, expressions, textes, ...) via l'activation de concepts. Pour la construction des vecteurs, nous avons pris deux hypothèses principales: l'*automatisation* de la création de la base lexicale vectorielle par apprentissage à partir d'informations extraites de diverses sources (dictionnaires à usage humain, liste de

*Ces recherches ont été effectuées dans le cadre de mon DEA Informatique au LIRMM, le LGI2P étant l'établissement qui finance actuellement mes recherches.

synonymes, ...), et un *apprentissage multi-source* afin de palier le bruit définitoire (par exemple les problèmes dus au métalangage comme dans la définition d'‹*aboyer*›, *crier* ***en parlant*** *du chien*). Chaque dictionnaire proposant un découpage des sens pour une entrée, nous avons mis en place une procédure qui associe à un sens un ensemble de définitions. Ce groupe de définitions est désigné par un identifiant numérique utilisé lors d'un processus de désambiguïsation. L'utilisateur, humain ou machine, doit donc connaître le codage pour pouvoir réassocier à un terme et un identifiant le sens correspondant. Chaque utilisateur doit dès lors posséder les sources lexicales du ou des désambiguïseurs auxquels ils font appel. Superviser une désambiguïsation manuellement demande systématiquement de consulter la source pour chaque terme polysémique. Nous proposons dans cet article de nommer un sens par un terme de la langue. Tout agent possédant une compétence linguistique doit être capable de retrouver le sens uniquement à partir du terme annoté (couple (terme, annotation)). Une telle procédure offre l'avantage d'une bonne interopérabilité; différents désambiguïseur peuvent proposer leurs résultats à différents clients (traducteur, indexeur, ...) et ces derniers peuvent faire appel à de multiples désambiguïseurs pour pallier les lacunes de certains. Nous présentons, dans un premier temps, le modèle des vecteurs conceptuels après lequel nous détaillons la procédure d'annotation et les différents aspects formels qui lui sont attachés.

1 Modèle vectoriel pour la sémantique lexicale

Le modèle vectoriel n'est pas récent, puisqu'il a été a introduit par Salton en informatique documentaire [14]. Sa réhabilitation dans les recherches en TALN est en revanche relativement récente, car elle a été essentiellement motivée par la mise à disposition des chercheurs de grandes bases de textes grâce au Web en particulier, alors que précédemment, ces recherches passaient par des phases ardues de constitution de corpus d'expérience. L'approche que nous avons s'inspire de la version de 1983 du modèle vectoriel de Salton [15], mais elle en diffère en ce que nous faisons l'hypothèse qu'il existe un jeu de concepts prédéterminé qui peut jouer le rôle d'ensemble générateur et que ce jeu est celui défini par les lexicologues quand ils réalisent un thésaurus [1]. Les concepts de cet ensemble sont par définition interdépendants: la famille considérée n'est pas libre et ne constitue pas une base vectorielle proprement dite. Cette interdépendance est aussi attestée dans un modèle comme LSA [2] qui non seulement la reconnaît mais aussi l'exploite.

Le modèle vectoriel a été appliqué par Salton à l'indexation et à la recherche d'information textuelle en 1988 [16]. Si ce dernier utilisait une analyse de surface par mots-clés pour alimenter ses vecteurs, notre démarche s'en distingue nettement: elle se base explicitement pour son calcul sur la géométrie et les variables morphosyntaxiques des arbres d'analyse structurelle issus du texte. D'une façon générale, les documents sont traités indépendamment les uns des autres, alors que dans LSA, le traitement se fait de manière liée. De plus nous mettons l'accent sur la sélection lexicale en contexte (voir Bourrigault 1993 *op. cit.*) alors que des travaux comme celui de [13] font un usage exclusif de taxinomies.

1.1 Avantages du modèle vectoriel pour la représentation du sens

Le modèle de vecteurs conceptuels s'appuie sur la projection dans un modèle mathématique de la notion linguistique de champ sémantique. Tout terme (lexie) et tout concept est projetable sur les vecteurs de la famille génératrice, et est donc représenté par un vecteur *conceptuel*. Mieux encore, on peut calculer le thème de tout segment de texte tel que documents, paragraphes, syntagmes, etc. sous forme de vecteur conceptuel: c'est le *sens* du segment en question [6]. Un vecteur correspond donc à une combinaison linéaire d'autres idées, termes, sens, et donc une vision de réseau et d'interdépendance est indissociable de ce modèle. Cette représentation homogène du sens, quelle que soit la granularité, est très avantageuse pour la classification des textes, l'indexation et la recherche évoluée d'information.

De plus, la représentation vectorielle ne fait aucune hypothèse *a priori* sur les relations conceptuelles. L'ontologie de départ mise à part, on ne se fonde sur aucune relation casuelle pour dériver du sens, et on n'inclut aucune contrainte sémantique. C'est un modèle purement calculatoire qui donne une *image* sémantique instantanée dans un état donné du dictionnaire conceptuel. Ce dernier est en apprentissage permanent, avec augmentation des définitions dès lors qu'une nouvelle source lexicologique électronique est disponible.

1.2 Relations sémantiques induites

Dans cet espace vectoriel conceptuel, on sait définir une notion de proximité sémantique en calculant une distance angulaire entre vecteurs (section 2.2). Cela signifie que l'on a une représentation de sens *proches*, sans pour autant valoriser correctement cette proximité. Le formalisme développé ci-après amène quelques remarques. On ne sait pas bien encore décliner cette proximité en relation d'hyperonymie ou d'hyponymie, qui sont caractéristiques des ontologies. En revanche, on arrive assez bien à mettre en valeur des relations transversales telles que la synonymie et l'antonymie [18] et qui sont très utiles lorsqu'il s'agit justement de faire émerger une microstructuration. Les paragraphes suivants définissent les propriétés générales de ces fonctions lexicales telles que nous les avons expérimentées dans [8] (*op. cit.*) et [17] (*op. cit.*).

2 Le modèle des vecteurs conceptuels

2.1 Principe

Soit $\mathcal{C}$ un ensemble fini de n concepts. Un vecteur conceptuel V est une combinaison linéaire des éléments c_i de $\mathcal{C}$. Pour un sens A, le vecteur V_A est la description (en extension) des activations des concepts de $\mathcal{C}$.Par exemple, les sens de ‹*ranger*› et de ‹*couper*› peuvent être projetés sur les concepts suivant (les *CONCEPT*[intensité] étant ordonnés par intensité[1] décroissante):

V_{ranger} = (*CHANGEMENT*[0.84], *VARIATION*[0.83], *ÉVOLUTION*[0.82], *ORDRE*[0.77], *SITUATION*[0.76], *STRUCTURE*[0.76], *RANG*[0.76] ...).

V_{couper} = (*JEU*[0.8], *LIQUIDE*[0.8], *CROIX*[0.79], *PARTIE*[0.78] *MÉLANGE*[0.78] *FRACTION*[0.75] *SUPPLICE*[0.75] *BLESSURE*[0.75] *BOISSON*[0.74] ...).

La description du processus d'apprentissage calculant les valeurs respectives des intensités pour chaque coordonnées d'un vecteur est exposé dans [7]. Il est clair que, pour des vecteurs denses (ayant très peu de coordonnées nulles), l'énumération des concepts activés est vite fastidieuse et surtout difficile à évaluer. On préférera en général procéder par sélection de termes thématiquement proches. Par exemple, les termes proches (et ordonnés par distance thématique décroissante) des mots ‹*ranger*› et ‹*couper*› sont:

‹***ranger***›: ‹*trier*›, ‹*cataloguer*›, ‹*sélectionner*›, ‹*classer*›, ‹*distribuer*›, ‹*grouper*›, ‹*ordonner*›, ‹*répartir*›, ‹*aligner*›, ‹*caser*›, ‹*arranger*›, ‹*nettoyer*›, ‹*distribuer*›, ‹*démêler*›, ‹*ajuster*› ...

‹***couper***›: ‹*cisailler*›, ‹*émincer*›, ‹*scier*›, ‹*tronçonner*› , ‹*ébarber*›, ‹*entrecouper*›, ‹*baptiser*›, ‹*recouper*›, ‹*sectionner*›, ‹*bêcher*›, ‹*hongrer*›, ‹*essoriller*›, ‹*rogner*›, ‹*égorger*›, ‹*écimer*›, ...

En pratique, plus $\mathcal{C}$ est grand, plus fines seront les descriptions de sens offertes par les vecteurs, mais plus leur manipulation informatique peut être lourde, surtout si l'on traite beaucoup de données. On rappelle que dans nos expérimentations sur le lexique général, $\dim(\mathcal{C}) = 873$, ce qui correspond au niveau 4 des concepts définis dans (Larousse, *op. cit.*) La construction d'un lexique conceptuel (ensemble de triplets *(mot, variables morphologiques, vecteur)*) est réalisée automatiquement à partir de corpora (de définitions, de thésaurii, etc. (Lafourcade *op. cit.*)). Au moment de l'écriture de cet article, le corpus du français représente environ $496\,658$ définitions correspondants à $136\,316$ mots vedettes (pour $62\,896$ mots monosémiques et $73\,420$ mots polysémiques – pour ces derniers le nombre moyen de définitions, certaines éventuellement redondantes, étant de 4.908).

[1] L'intensité est une valeur positive comprise entre 0 et 1 permettant de relativiser la présence de chaque concept. Une normalisation des vecteurs est effectuée, seul les angles sont utilisés et non les normes de vecteurs.

2.2 Distance angulaire

Il est souhaitable de pouvoir mesurer la proximité entre les sens représentés par deux vecteurs (et donc celle de leur mot associé). Soit $Sim(X, Y)$ la mesure de *similarité*, utilisée habituellement en recherche d'informations, entre deux vecteurs définie selon la formule **(1)** ci-dessous (avec "·" étant le produit scalaire). On notera que l'on suppose ici que les composantes des vecteurs sont toujours positives ou nulles (ce qui n'est pas nécessairement le cas). Enfin, nous définissons une fonction de *distance angulaire* D_A entre deux vecteurs X et Y selon la formule **(2)**.

$$\textbf{(1)} \qquad Sim(X,Y) = \cos(X,Y) = \frac{X \cdot Y}{\|X\| \times \|Y\|}$$

$$\textbf{(2)} \qquad D_A(X,Y) = \arccos(Sim(X,Y))$$

Intuitivement, cette fonction constitue une évaluation de la *proximité thématique* et est en pratique la mesure de l'angle formé par les deux vecteurs. On considérera, en général, que pour une distance $D_A(X, Y) \leq \pi/4$ (soit environ $0,78$ radian ou encore 45 degrés), X et Y sont sémantiquement proches et partagent des concepts. Pour $D_A(X, Y) \geq \pi/4$, la proximité sémantique de A et B sera considérée comme faible. Aux alentours de $\pi/2$ (soit environ $1,57$ radians ou 90 degrés), les sens sont sans rapport. La synonymie (dans son acception la plus large) est incluse dans la proximité thématique, cependant elle exige, de plus, la concordance des catégories morphosyntaxiques. L'inverse n'est évidemment pas vrai.
La distance angulaire est une vraie distance (contrairement à la mesure de similarité) et elle vérifie les propriétés de réflexivité **(3)**, symétrie **(4)** et inégalité triangulaire **(5)** (qui peut jouer un rôle de pseudo-transitivité):

$$\textbf{(3)} \qquad D_A(X,X) = 0$$

$$\textbf{(4)} \qquad D_A(X,Y) = D_A(Y,X)$$

$$\textbf{(5)} \qquad D_A(X,Y) + D_A(Y,Z) \geq D_A(X,Z)$$

Par définition, nous posons: $D_A(\vec{\mathbf{0}}, \vec{\mathbf{0}}) = 0$ et $D_A(X, \vec{\mathbf{0}}) = \pi/2$ pour tout X avec $\vec{\mathbf{0}}$ dénotant le vecteur nul[2]. On considérera, en toute généralité, l'extension du domaine image de D_A à $[0, \pi]$ afin de comparer des vecteurs ayant des composantes négatives. Cette généralisation ne change pas les propriétés de D_A. On remarquera, de plus, que la distance angulaire est insensible à la norme des vecteurs (α et β étant des scalaires):

$$D_A(\alpha X, \beta Y) = D_A(X,Y) \quad \text{avec} \quad \alpha\beta > 0$$

$$D_A(\alpha X, \beta Y) = \pi - D_A(X,Y) \quad \text{avec} \quad \alpha\beta < 0$$

[2] Le vecteur n'est sans doute pas représenté par un mot de la langue. Il s'agit d'une idée qui n'active... aucun concept ! C'est l'idée vide.

Par exemple[3] dans le tableau qui suit, nous avons les distances angulaires (en radian) entre les vecteurs de plusieurs termes. Le tableau est symétrique (à cause de la symétrie de D_A) et la diagonale est toujours égale à 0 (à cause de la réflexivité de D_A). On remarquera qu'une valeur prend toute sa signification relativement à une autre. En particulier, il est satisfaisant d'avoir: (a) $d_1 \leq d_3$ et $d_2 \leq d_3$ ce qui correspond bien au fait que ‹*trier*› et ‹*ordonner*› d'une part, et ‹*trier*› et ‹*choisir*› sont "plus synonymes" que ‹*ordonner*› et ‹*choisir*›. On remarquera aussi que d_3 est supérieure à $\pi/4$, ce qui dénote un éloignement sémantique qui commence; (b) d_4 est la plus petite valeur de $D_A(\text{ranger}, Y)$ car les concepts *CLASSER* et *RÉPARTIR* sont relativement proches, et de plus ‹*ranger*› est par ailleurs polysémique (*CLASSER*, *RASSEMBLER* et *NETTOYER*) et seul *CLASSER* est présent dans le tableau.

$D_A(X,Y)$	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
(1) trier	0	0,517	0,662 d_1	0,611 d_2	0,551	0,441	0,462
(2) ranger		0	0,829	0,6	0,523	0,409 d_4	0,444
(3) choisir			0	0,848 d_3	0,77	0,796	0,758
(4) ordonner				0	0,595	0,523	0,519
(5) ventiler					0	0,471	0,391
(6) classer						0	0,36
(7) répartir							0

L'espace vectoriel conceptuel est muni de deux lois de composition interne: la somme (et son opération symétrique, la soustraction) et le produit terme à terme (on ne définit pas ici son opération symétrique) qui sont détaillées dans le prochain paragraphe.

3 Annotation sémantique, nommage de sens: définition

Nombres de systèmes actuels associent à chaque mot polysémique une annotation numérique [4], [22], par exemple:

"*Le chat*/**I.2**/ *mange*/**1**/ *la souris*/**II.1**/".

L'usage que nous proposons diffère des précédents, nous souhaitons retrouver des associations d'idées dans la langue et les utiliser pour désigner un sens. Nous associons à chaque terme polysémique d'un texte un annotateur qui est lui même un terme du lexique. Par exemple:

> *Chaussé*/**porter**/ de ses *bottes*/**chaussure**/ il *revenait*/ **déplacer**/ vers la grange et *apercevait*/**voir**/ les *bottes*/ **amas**/ de *foin*/**paille**/.

Le lexicographe qui dispose ainsi d'un terme auquel est associé une annotation est alors capable d'identifier plus facilement le sens désigné. Il ne s'agit plus de donner une définition comme annotateur [23] mais d'extraire un représentant. Par exemple, nous proposons d'annoter le terme ‹*botte*› par *botte*/**paille**/, *botte*/**chaussure**/, *botte*/**escrime**/ qui sont des formes plus intuitives et compréhensibles. On peut alors considérer que cette annotation est *hors-source*, elle fait sens sans source lexicale spécifique. Si le destinataire a une compétence linguistique, il lui sera possible de réassocier la botte et la chaussure, ou la botte et la réunion de végétaux.

3.1 Définition formelle

L'annotation correspond à une fonction bijective qui associe à un terme et un sens un couple *(terme, annotation)*. Soit D le dictionnaire, M un terme du dictionnaire, s_i un sens de M tel que $s_i = M$ et A_i un ensemble d'annotateurs pour s_i et soit f la fonction d'annotation:

$$\forall M \in D, \forall s_i \in M$$

$$f : s_i \longrightarrow A_i$$

$$A_i \subset D$$

telle que:

$$\forall A_i = f(s_i), \quad \forall A_j = f(s_j), \quad i \neq j,$$

$$A_i \cap A_j = \emptyset \quad et \quad A_i, A_j \neq \emptyset$$

Quel que soit le terme $M \in D$, on partitionne le lexique pour obtenir pour chaque sens $s_i \in M$ un ensemble (non vide) d'annotateurs A_i. Cette fonction doit permettre de réassocier à tout couple (a, M) où a est un élément de A_i le sens s_i correspondant. Cette fonction admet donc une bijection réciproque f^{-1} :

$$\forall M \in D, \qquad \forall A_i \subset D,$$

$$f^{-1} : \forall a \in A_i \qquad (a, M) \longrightarrow s_i$$

On souhaite ainsi insérer dans un texte des balises contenant un terme discriminant pour chaque terme polysémique désambiguïsé. La fonction d'annotation précédente propose pour un sens donné d'un terme un ensemble de candidats. Il importe donc d'évaluer les différents candidats et de les classer par ordre d'intérêt tout en accordant une souplesse suivant l'utilisation qui sera faite de l'annotation (usage humain, interopérabilité de systèmes automatiques, ...).

3.2 Propriétés remarquables

Indépendance aux dictionnaires
Dans le cas où on ne souhaite pas diffuser le dictionnaire source (droits d'auteur, volume trop important, ...) une annotation par des termes de la langue peut permettre au client de retrouver le bon sens sans la source. Si dans plusieurs dictionnaires un terme est associé à un autre, il est fortement probable de retrouver la même relation à l'aide d'autres sources. Différents annotateurs n'offriront pas la même indépendance aux sources, il est donc important d'évaluer cette propriété et de classer les annotateurs en conséquence.

Une interface homme-machine
L'annotation est un outil précieux pour le lexicographe qui supervise une désambiguïsation. Elle met en évidence de façon simple

[3] Tous les exemples de cet article sont issus de <http://www.lirmm.fr/~lafourca>

le sens qu'il faut attribuer à un terme, sans devoir constamment se référer à un dictionnaire donné et en assimiler la définition. Le coût cognitif est défini comme l'effort que doit faire l'agent humain pour transformer une perception, une information en une connaissance exploitable [12]. Dans le cas de l'annotation sémantique à usage humain, il est déterminant de minimiser ce coût et pour cela de prendre en compte d'autres critères. On favorisera ainsi des candidats qui ont un usage proche en utilisant des notions de fréquence, de co-occurrence terme/annotation, mais aussi des informations fournies par les dictionnaires comme la morphologie ou l'usage (contexte ou domaine, sens figuré ou soutenu, ...).

Interopérabilité

L'indépendance aux sources facilite la coopération entre différents systèmes. Dans cette perspective, l'annotation sémantique représente à la fois un *pivot* de communication entre systèmes, mais aussi une interface permettant de diviser le traitement de la langue en *deux couches*.

Le rôle de pivot est celui que l'on recherche dans des évaluations comme SENSEVAL[4]. Une annotation commune à tous les systèmes est donnée et permet de les comparer. Cette approche apporte un intérêt nouveau: l'utilisation conjointe de multiples désambiguïsations. Dans ces compétitions, les différents systèmes sont imparfaits, mais les erreurs produites sont fréquemment disjointes. Nous proposons une approche multi-critère qui utilise un médiateur pour déterminer son résultat. Un client d'une désambiguïsation demande un résultat à chaque système, les compare. Il peut d'une part mettre en évidence les contradictions, d'autre part décider d'adopter ou non un résultat. S'il sélectionne un résultat, il lui est possible d'attribuer une *valeur de confiance* dans le résultat.

D'autre part, un problème actuel des systèmes de désambiguïsation est l'apprentissage. En effet, il est difficile pour un désambiguïseur d'évaluer automatiquement s'il fait une erreur ou non, et par conséquent d'analyser dans quelles conditions il effectue ses erreurs. Avec la notion de pivot précédemment décrite, un tel système peut en permanence comparer ses résultats et détecter des contradictions avec les autres systèmes. On peut ainsi proposer un *dialogue* entre systèmes, dans le sens ou chacun va remettre en cause ses résultats. Il va juger de sa confiance et reproposer un résultat. L'objectif est d'atteindre un accord global par itération, et un apprentissage par coopération.

Toutes ces coopérations se passent de l'utilisation d'un protocole de communication complexe, et simplement partagent un résultat qui est un texte de la langue annoté par des termes de la langue. Le protocole provient de la compétence linguistique commune à tous les agents. L'annotation joue aussi un rôle d'interface entre deux couches, l'une en amont de la désambiguïsation, l'autre en aval.

Enfin, si dans le cas de l'interopérabilité, la notion de coût cognitif apparaît moins primordiale, elle n'en reste pas moins importante, nous adressant à des systèmes supposés ayant une compétence linguistique. En effet, de nombreux modèles de représentation du sens en traitement automatique des langues se basent sur une approche statistique sur le distribution des mots, qu'elle soit simple ou évoluée, dans un corpus comme le Web ou dans un dictionnaire par exemple. Or, l'homme quand il rédige de tels documents imprègne dans cette conception de reflet de sa pensée. Donc le coût cognitif prend aussi un sens quand on s'intéresse à un systèmes automatique qui base ses connaissances sur des données humaines.

Augmentation lexicale

En anglais, le terme ‹*giblets*› signifie ‹*abats de volaille*›, le terme ‹*offal*› signifie ‹*abats de porc ou de boeuf*›. En français, ces deux sens ne sont lexicalisés que par l'hyperonyme ‹*abats*›. Lors d'une traduction, il est nécessaire de pouvoir traduire le terme anglais par une locution en français, et réciproquement, le lexique français doit pouvoir générer le sens anglais et l'associer à un terme pour éviter de générer pour ‹*abats de volaille*› la traduction ‹*giblets of fowl*›. L'annotation est adaptée à ce problème car en définitive, elle correspond à une extraction automatique de gloses et possède ainsi le pouvoir génératif nécessaire [5].

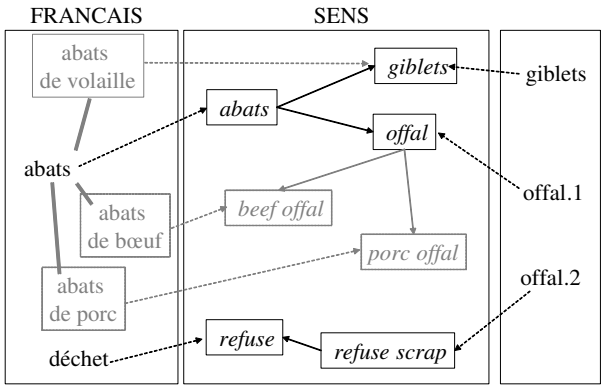

Figure 1: Ce schéma présente le principe de l'augmentation lexicale. Une base multilingue possède des acceptions, lexicalisées dans certaines langues mais non dans d'autres. On souhaite créer de nouvelles locutions dans les langues qui ne possèdent pas de terme pour désigner le sens.

4 Comportement de l'utilisateur

Nous l'avons vu, nous souhaitons annoter avec un terme du lexique, dans un but notamment d'interface homme-système ou système-système. Dans les deux cas, nous souhaitons des résultats comparables à ceux obtenus par le comportement de l'utilisateur. En effet, s'il s'agit d'interopérabilité entre processus automatiques, le pivot provient de l'hypothèse que les différents agents vont posséder une compétence linguistique. Nous avons donc procédé à une enquête sur le Web[5] afin d'identifier le comportement de l'utilisateur dans le choix d'un terme pour nommer un sens. Le sujet de l'étude se voyait proposer une liste de termes et pour chacun différents sens (une liste non exhaustive). Quand cela était nécessaire, nous précisions une définition succincte. L'utilisateur devait alors sélectionner un terme pour nommer le sens, et si aucun ne lui convenait, il pouvait en proposer un. Le nombre de sens à nommer était de 38 pour 13 mots, tandis que 134 réponses ont été obtenues, uniquement par volontariat. Le questionnaire avait une durée moyenne de l'ordre de 15 à 20 minutes, et il était possible à tout moment d'interrompre l'étude en ne renvoyant qu'un résultat partiel.

4.1 Qualité des réponses

Les taux d'abstention n'ont pas fortement augmenté durant l'enquête, malgré sa durée et la non rémunération. D'autre part,

[4] *http://www.itri.brighton.ac.uk/events/senseval/*

[5] Le détail et les résultats de cette enquête sont disponibles à l'adresse *http://www.lgi2p.ema.fr/∼jalabert/recherche/annot-hs/sondage.html*

la consigne pourtant compliquée a été très bien comprise et les réponses ont été faites avec beaucoup de sérieux. Des erreurs sont présentes, mais témoignent plus de la difficulté pour l'Homme de désambiguïser, c'est à dire de sélectionner un sens parmi tous ceux existants et ne relevant pas de l'incompréhension de la consigne. Pendant nos interactions langagières, le mécanisme que nous utilisons n'est pas un choix dans une liste exhaustive de sens pour chaque terme. Les ambiguïtés se lèvent dans le contexte et posent rarement de problème. Il est extrêmement difficile pour une personne à un instant donné, sans source lexicale disponible, de pouvoir répertorier tous les sens existants pour des termes comme ‹*maison*›, ‹*courant*›, ‹*carte*›, ‹*défendre*›, ... Lorsque nous faisons appel à une telle démarche chez le sujet de l'expérience, cette difficulté influe donc sur le résultat.

4.2 Adéquation de l'approche multi-critère

Le résultat de cette enquête montre avant tout l'adéquation de l'approche multi-critère avec le problème. En effet, il n'y a pas de méthode qui soit adaptée en toutes situations et pour toutes personnes. Cependant, on retrouve trois comportements principaux:

L'association de termes est souvent utilisée, avec par exemple ‹*botte de paille*›, ‹*botte de foin*›, etc.

La substituabilité de termes est le deuxième motif principalement utilisé. Complémentaire avec l'association de terme, l'utilisateur sélectionne des synonymes, en recherchant la plupart du temps un terme plus général (un hyperonyme). Par exemple, ‹*poisson*› se voit associer ‹*animal*› et non un hyponyme comme ‹*truite*›, ‹*requin*›, etc.

L'identification du contexte, du domaine est enfin le dernier comportement qui apparaît dans le cas où la personne ne connait pas le sens. Il choisit alors d'induire le contexte et de le nommer (‹*technique*›, ‹*mécanique*›, etc.).

En définitive, cette enquête témoigne de l'adéquation de l'approche multi-critère. Les synonymes sont souvent adaptés, et peuvent être extraits principalement de dictionnaires, les associations d'idées quand elles sont bien marquées peuvent être privilégiées et peuvent être obtenues par des analyses statistiques de corpus (collocation). Enfin, l'utilisation des vecteurs et de thésaurus permet de déterminer la thématique d'un segment textuel et d'identifier le domaine.

4.3 Importance de la génération flexionnelle

Les résultats obtenus montrent que *botte*/**secret**/ ne représente pas une association forte pour l'utilisateur, et que le coût cognitif est relativement important. Au contraire, *botte*/**secrète**/ sera beaucoup plus adapté. L'utilisation de flexions est très importante dans le cas de l'annotation de textes, où les verbes sont généralement conjugués. Il est donc important d'associer à la procédure d'annotation une méthode permettant d'accorder l'annotateur en fonction du mot annoté et de son contexte.

Ce besoin de retrouver les flexions s'exprime autant au niveau de l'interface homme-machine qu'à celui de l'interopérabilité de systèmes automatiques. Si pour le premier la motivation est évidente, pour le second nous avons fait l'hypothèse que le système est doté d'une compétence linguistique et que cette compétence lui permet de désambiguïser. Dans une vision plus *pratique*, les systèmes basés sur une modélisation distributionnelle de la langue ont bien plus de chances de retrouver la collocation ‹*botte secrète*› dans un texte que ‹*botte secret*›. De façon réciproque, dans une analyse statistique de corpus, la lemmatisation se montre tout aussi indispensable.

Les tableaux suivants résument les résultats obtenus avec Google[6] concernant la co-occurrence et la collocation de l'exemple précédent.

$$hit(\text{‹}botte\text{›}) = 53700$$

t_i	$hit(t_i)$	$hit(\text{‹}botte\text{›}, t_i)$	$Coocc(\text{‹}botte\text{›}, t_i)$
‹*secret*›	105 000	3110	$1,72.10^{-3}$
‹*secrète*›	160 000	4810	$2,39.10^{-3}$
‹*secret*› ou ‹*secrète*›	220 000	6890	$4,02.10^{-3}$

avec $Coocc(t_1, t_2) = \frac{hit(t_1,t_2)}{hit(t_1).hit(t_2)}$

Table 1: Résultats de la co-occurrence sur le terme ‹*botte*› et certains de ses candidats

t_i	$hit(t_i)$	*fréquence*
‹*botte secrète*›	1590	2,96 %
‹*botte secret*›	5	9,3 . 10^{-3} %

Table 2: Fréquence des collocations ‹*botte secret*› et ‹*botte secrète*› parmi les occurrences de ‹*botte*›.

4.4 Complément d'enquête

L'enquête que nous avons menée a conforté notre approche du problème. Cependant, elle a révélé d'autre part que d'autres études seraient nécessaires. En effet, les termes contenus dans les définitions des sens donnés ont fortement influé sur les réponses, résultat prévisible. Nous aimerions donc effectuer d'autres expériences avec des protocoles différents. Ainsi nous souhaiterions confronter l'utilisateur à la tâche exacte qu'il demande au systèmes, c'est à dire étant donné un texte, annoter chaque terme suivant une liste de mots fournis ou de façon totalement libre (aucune suggestion). Pour sélectionner le sens, le sujet devrait alors uniquement interpréter le contexte. De même, des expériences plus spécifiques à l'annotation de requête courtes en recherche d'information et à la sélection de termes traduits dans une langue non maitrisée.

De plus, [21] présente une étude témoignant du désaccord entre les hommes pour annoter un document: pour un texte donné, différents linguistes ne sélectionnent pas forcément les mêmes sens, problème d'autant plus fréquent dans le cas de continuité des sens et de verbes supports par exemple, mais aussi dans le cas de la multisémie (sélection multiple de sens). Il remet ainsi en question l'utilisation même de l'annotation pour évaluer des désambiguïseurs. Nous souhaiterions compléter cette étude en évaluant, dans un texte annoté pour des termes possédant une polysémie plus marquée, si l'utilisateur accepte ou non l'annotation. Il serait important de mettre l'utilisateur dans une situation réelle, dans le

[6] *http://www.google.fr*

cadre d'un moteur de recherche ou un traducteur automatique par exemple. Cela donnerait la possibilité de mesurer la validité de l'annotation chez l'utilisateur, mais aussi de pouvoir comparer la durée qui lui est nécessaire pour réassocier à un terme le sens correct en fonction de l'annotateur, et de comparer ainsi différentes méthodes d'annotation.
Enfin, ces différentes réflexions montrent un besoin d'ouverture à l'expertise d'autre disciplines. Ainsi, pour ces études, une collaboration avec des psychologues, psychosociologues et ergonomes serait nécessaire, tout comme les sciences du vivant pourraient contribuer dans le futur en apportant une meilleur connaissance du fonctionnement du cerveau et du langage.

5 Procédure d'annotation

Notre procédure d'annotation repose sur une approche multicritère traitant de différentes façons de multiples sources lexicales, mais utilisant aussi d'autres représentations dont les vecteurs conceptuels. Pour des raisons pratiques (coût calculatoire élevé), nous utilisons dans un premier temps les sources lexicales pour extraire les candidats. Un filtre est ensuite appliqué afin d'optimiser la probabilité de réassocier le bon sens pour un couple (terme, annotation) donné. Enfin, nous souhaitons ordonner les candidats pour déterminer lequel est le plus adéquat pour une application donnée en recherchant un compromis entre la *capacité à désambiguïser un sens* et la *capacité à s'associer avec*.

5.1 Extraction de candidats

L'extraction de polysèmes s'organise autour de deux types de sources lexicales. Le premier est une information provenant de dictionnaires à usage humain habituels, dans laquelle on peut isoler différents sens pour un même terme. Le second que nous nommons *sacs de polysèmes* correspond aux thésaurus, listes de synonymes, d'antonymes, etc où les mots sont donnés sans distinction de sens (par exemple[7]: ‹*botte*› → *assemblage, attaque, balle, boots, bottée, bottelée, bottelette, bottillon, bottine, bouchon,* ...).

Définitions rattachées à un sens
Ce premier type d'extracteur s'applique à un ensemble de dictionnaires traditionnels, analysant les différentes définitions et proposant tous les mots contenus comme candidats. Nous utilisons pour cela SYGMART [8], qui est une plateforme de conception d'outils d'analyse textuelle disposant notamment d'un analyseur morphosyntaxique du français. Ce dernier calcule à partir d'une phrase donnée un arbre morphosyntaxique. On obtient alors des informations complètes sur la morphologie des termes de la phrase ainsi que leur fonction. Les les articles, les pronoms, le métalangage (*famille de, du latin,...*) ainsi filtrés car ils participent moins à la thématique de la définition. L'analyse morphosyntaxique permet aussi de gérer les appositions de compléments fréquentes dans les définitions. Les définitions étant décrites principalement en *genre et différence*, on donne la plus forte pondération aux premiers termes de la définition, au mot-tête s'il est présent. Les premiers termes sont souvent des candidats pertinents.
D'autre part, certains dictionnaires électroniques proposent des données dans un langage semi-structuré (SGML, XML, ...) dont on peut extraire des informations pertinentes. Par exemple, certaines encyclopédies détachent pour les noms propres un résumé très adapté à notre approche (‹*Napoléon Bonaparte*› - *empereur français*).

Sacs de polysèmes
Ces extracteurs s'appliquent à des sources comme une liste de synonymes [11] , concepts du thésaurus, des dictionnaires de co-occurrence, ... Les informations obtenues correspondent à un terme et non à un sens. La nature de ces sources est souvent pertinente, mais on accorde moins de confiance à ces candidats.

5.2 Validation d'un candidat

Après avoir extrait de nombreux candidats, il est nécessaire de les évaluer, mais une étape préliminaire consiste à appliquer un filtre qui valide ou non des candidats. Comme nous l'avons expliqué précédemment, nous souhaitons obtenir une fonction admettant une réciproque pour associer à un couple (a_i, M) un sens $s_i \in M$. L'objectif de la procédure de validation est simplement d'ôter tout candidat qui ne satisfait pas cette condition. Elle se formalise de la façon suivante: un annotateur a est valide si et seulement si:

$$\forall s_j \neq s_i, \qquad D_A(a, s_i) \leq D_A(a, s_j)$$

La première étape de la validation correspond à la validation des candidats associés à des sens: si un candidat est plus proche d'un autre sens que celui qu'il annote, alors il est éliminé. D'autre part, si un candidat est extrait d'un sac de polysèmes, la seconde étape lui associe le sens le plus proche. Enfin, tous les candidats présents dans une définitions appartenant à un autre sens sont considérés comme source d'ambiguïté et sont eux aussi éliminés. On peut ainsi détecter les termes qui ont été mal affectés, mais aussi les termes généraux qui sont présents dans de nombreuses définitions sans participer réellement à la thématique (le métalangage, les verbes supports, les termes du *vague*, ...).

5.3 Évaluation d'un candidat

L'évaluation est organisée sous trois angles. Le premier est l'utilisation d'une note d'extraction en fonction de la position structurelle du terme (dans l'arbre morphosyntaxique). Les définitions étant décrites en *genre et différence*, les premiers termes extraits sont fréquemment des hyperonymes ou caractéristiques apportant un bon compromis entre désambiguïsation et association d'idées. D'autre part, on évalue ces deux derniers points indépendamment. L'évaluation de la désambiguïsation correspond donc à trois critères:

- une note d'extraction,
- une formalisation de la notion de *capacité de désambiguïsation*,
- et enfin l'extraction de l'association entre termes dans l'usage.

Une note d'extraction des candidats
Durant la phase d'extraction des candidats, une évaluation est effectuée suivant: (1) *la source lexicale* (différents dictionnaires n'offrent pas la même qualité de résultat), (2) *le type d'extraction* (les concepts les plus activés ou les domaines et usages qui ne sont pas le plus souvent les meilleurs) et enfin (3) *l'arbre morphosyntaxique* fournit des informations comme la position et la fonction du terme dans la phrase qui sont particulièrement précieuses.

Capacité de désambiguïsation
L'objectif est d'obtenir une annotation pour la désambiguïsation.

[7] Extrait du dictionnaire des synonymes de l'Université de Caen - *http://elsap1.unicaen.fr/dicosyn.html*
[8] développé par Jacques Chauché: *http://www.lirmm.fr/∼chauche/Pr%E9sentationSygmart.html*

Il s'agit ici d'évaluer dans quelle mesure un terme permet de distinguer un sens de tous ses concurrents. Pour cela, nous avons introduit une notion de marge de désambiguïsation qui tente de représenter la probabilité d'erreur d'un système pour réassocier le bon sens.

La marge de désambiguïsation absolue
Elle représente l'intervalle dans lequel la fonction réciproque n'associe pas l'annotateur à un mauvais sens. Plus cette marge est importante, plus les probabilités sont bonnes de ne pas réassocier un mauvais sens dans une autre base lexicale. Soit a_1 l'annotateur de s_1 et $s_2 \neq s_1$ le second sens le plus proche de a_1. Alors la marge absolue est:

$$M_A(a_1, s_1) \quad = \quad | D_A(a_1, s_2) - D_A(a_1, s_1) |$$

La marge de désambiguïsation relative
. Deux candidats ayant une même marge absolue ne sont pas forcément équivalents. Si un des termes est plus proche du sens à annoter, la marge est proportionnellement plus importante. La marge relative suivante permet de prendre en compte cette notion de proportionnalité:

$$M_R(a_1, s) \quad = \quad \frac{M_A(a_1, s)}{d_1}$$

Le risque de non-sens
Si deux termes ont une probabilité d'erreur comparable pour réassocier le sens, il est important de tenir compte de la gravité de l'erreur. Les deux schémas suivants (Figures 3 & 4) présentent les associations entre les différents sens des candidats ‹*voilier*› et ‹*vaisseau*› pour annoter le sens de ‹*frégate*› correspondant au navire ancien. Si une erreur est faite pour l'annotation ‹*voilier*›, on va interpréter le sens de l'oiseau au lieu du précédent. Pour ‹*vaisseau*›, on associera celui du ‹*bâtiment de guerre*›. L'utilisation de ‹*voilier*› risque donc de produire un contre sens alors que dans l'autre cas, on continuera à parler d'un navire. Lors d'une traduction en anglais, une erreur sera produite dans le premier cas, car le sens de l'oiseau de traduit par ‹*frigate bird*›, alors que les deux navires correspondent au terme ‹*frigate*›.
Le risque de non-sens est:

$$R_{NS}(a_1, s) \quad = \quad \frac{d_3}{M_R(a_1, s)}$$

où d_3 est la distance qui sépare les deux sens pour lesquels il y a ambiguïté. Le risque de non-sens ajoute donc la possibilité d'obtenir un compromis en la probabilité de faire une erreur de désambiguïsation et la gravité de cette erreur.

Association terme - candidat

Pour sélectionner un terme réellement associé, il est déterminant de tenir compte de l'usage de l'annotateur et de l'annoté. Pour cela, plusieurs évaluateurs sont présents dans notre procédure:
(1) *La fréquence d'usage de l'annotateur*: supposons que ‹*mouche*› reçoive comme candidat *mouche*/**drosophile**/ (pour le sens de l'insecte). Un annotateur très rare risque de ne pas être connu du client, qu'il soit humain ou automatique, un annotateur trop fréquent risque de ne pas faire sens (*cuisine*/**faire**/). De même, suivant l'utilisation, on peut souhaiter dans l'ordre de préférences un hyperonyme, un hyponyme ou un co-hyponyme. La fréquence est un indice: un terme bien moins fréquent qu'un autre a plus de chance d'être un hyponyme qu'un hyperonyme.

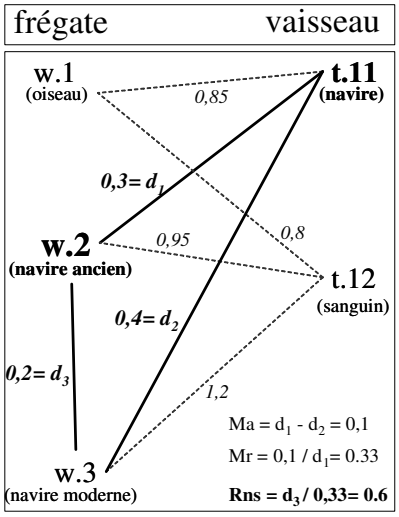

Figure 2: *Marges de désambiguïsation* et *risque de non-sens* pour le candidat *frégate*/**vaisseau**/

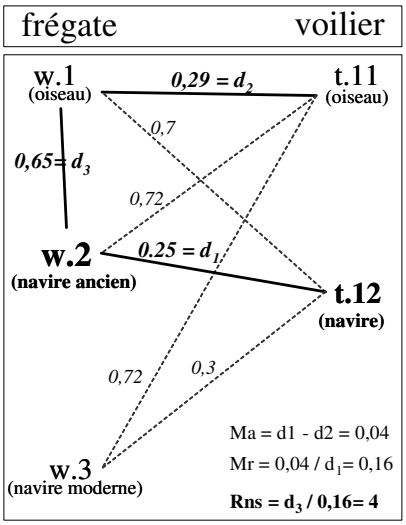

Figure 3: *Marges de désambiguïsation* et *risque de non-sens* pour le candidat *frégate*/**voilier**/. ‹*voilier*› est polysème, il s'agit non seulement d'un type de bateau, mais aussi d'une espèce d'oiseau.

(2) *La catégorie grammaticale*: le choix d'utiliser des termes de même nature grammaticale réduit le coût cognitif, technique qu'utilisent depuis longtemps les dictionnaires pour définir un terme (‹*déplanter*› → ‹*enlever*›; ‹*rimer*› → ‹*constituer une rime*›; ‹*table*› → ‹*meuble*› ...).
(3) *Marqueurs d'usage*: deux utilisateurs différents n'auront pas les mêmes associations d'idées entre différents mots. Par exemple, le terme ‹*police*› peut signifier le contrat d'assurance ou l'autorité judiciaire. Dans ce deuxième sens, l'annotation ‹*agent*› ou ‹*poulet*› n'induira pas le même comportement chez le lecteur. Celui-ci risque de ne pas retrouver une association pourtant évidente chez un autre. La co-occurrence permet de confirmer l'association d'idée entre deux termes. De plus, l'exemple précédent montre qu'une base de connaissance distributionnelle tenant

compte des contextes et usages offrira de meilleurs résultats. Un site personnel et un livre ne s'adresseront pas aux même lecteurs, tout comme un article de presse et un article scientifique. Les définitions possèdent certains marqueurs qui permettent d'associer ou dissocier des termes:

- des marqueurs d'usage (‹*péjoratif*›, ‹*vieux*› ou ‹*ancien*›, ...),
- des marqueurs de niveau de langue (‹*poétique*›, ‹*familier*›, ...),
- des marqueurs de domaine (‹*médecine*›, ‹*zoologie*›, ...).

Conclusion et perspectives

L'annotation telle que nous venons de la présenter est un outil important dans le cadre de nos recherches. C'est une interface homme-machine permettant une évaluation rapide et efficace pour le superviseur d'un processus de désambiguïsation dans le cadre de la traduction, ou dans notre cas, une aide de l'analyse de définitions pour produire un vecteur conceptuel. Elle est, d'autre part, une nouvelle forme d'interfaçage entres de multiples systèmes. Les résultats actuels sont encourageants, mais de nombreuses voies restent à explorer. Nous projetons ainsi dans un avenir proche d'étudier plus précisément les comportement de l'utilisateur et d'expérimenter l'interfaçage entre différents systèmes automatiques dans un cadre multilingue. Enfin, cette analyse des associations d'idées va devenir une nouvelle source lexicale pour la base vectorielle. L'objectif est d'améliorer la base de connaissance, et de ce fait notre propre processus (phénomène de la double boucle [20]. Des améliorations seraient finalement nécessaires par l'ajout d'une mémoire associative (‹*Tintin*› → ‹*Milou*›) pour l'extraction et l'évaluation des candidats, mais aussi l'utilisation des fonctions lexicales (métonymie, méronymie, holonymie, ...).

References

[1] Chauché J., *Détermination sémantique en analyse structurelle: une expérience basée sur une définition de distance.* TAL Information, 31/1, pp 17-24, 1990.

[2] Deerwester S., Dumais S., Landauer T., Furnas G., Harshman R., *Indexing by latent semantic analysis*, Journal of the American Society of Information Science, 416(6), p. 391-407, 1990.

[3] Hachette. *Dictionnaire Hachette Encyclopédique.* Hachette, ISBN 2.01.280477.2, version en ligne: http://www.encyclopedie-hachette.com

[4] Ide N., Erjavec T., Tufis D. *Automatic Sense Tagging Using Parallel Corpora* of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, 83-9.

[5] Jalabert F., Lafourcade M. *From word sense tagging to vocabulary augmentation in Papillon* Papillon 2003, Sapporo, Japon, Juillet 2003.

[6] Lafourcade M. et E. Sandford *Analyse et désambiguïsation lexicale par vecteurs sémantiques* TALN'1999, Cargèse, France, Juillet 1999, pp 351-356.

[7] Lafourcade M., *Lexical sorting and lexical transfer by conceptual vectors*, First International Workshop on MultiMedia Annotation (MMA'2001), Tokyo, 6 p, January 2001.

[8] Lafourcade M., Prince V., *Synonymies et vecteurs conceptuels*, TALN 2001, Tours, p. 233-242, juillet 2001.

[9] Lafourcade M., Prince V., Schwab D. *Vecteurs conceptuels et structuration émergente de terminologies* Revue TAL Volume 43 - n 1/2002, pages 43 à 72

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

[11] Ploux S., Victorri B. *Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes* Revue TAL Volume 39, 1998, Numéro 1.

[12] Prince V. *Vers une informatique cognitive dans les organisations - Le rôle central du langage* Ed. Masson 1996.

[13] Resnik P., *Using Information contents to evaluate semantic similarity in a taxonomy*, IJCAI-95, 1995.

[14] Salton G. *Automatic Information Organisation and Retrieval* McGraw-Hill, New York 1968.

[15] Salton G., MacGill M.J., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[16] Salton G., *Term-Weighting Approaches in Automatic Text Retrieval*, McGraw-Hill computer science series, McGraw-Hill, vol. 24, 1988.

[17] Schwab D., *Vecteurs conceptuels et fonctions lexicales: application à l'antonymie*, Mémoire de DEA Informatique, 2001.

[18] Schwab D., Lafourcade M., V. Prince V., *Amélioration de la représentation sémantique lexicale par les vecteurs conceptuels: le rôle de l'antonymie*, JATD 2002, vol. 2, p. 701-712, 2002.

[19] Schwab D., Lafourcade M., Prince V. *Vers l'apprentissage automatique, pour et par les vecteurs conceptuels, de fonctions lexicales. - L'exemple de l'antonymie.* TALN'2002 Nancy, Juin 2002.

[20] Schwab D. *Société d'agents apprenants et sémantique lexicale: comment construire des vecteurs conceptuels à l'aide de la double boucle.* RECITAL 2003, Batz-sur-Mer, Juin 2003.

[21] Véronis J. *Sense tagging: does it make sense?* Corpus Linguistics'2001 Conference, Lancaster, U.K.

[22] Wiebe J. , Maples J. , Duan L. , Bruce R. *Experience in WordNet sense tagging in the Wall Street Journal*. In Proc. ANLP-97 Workshop, Tagging Text with Lexical Semantics: Why, What, and How? Association for Computational Linguistics SIGLEX, Washington, D.C., April 1997, pp. 8-11.

[23] Wilks Y., Stevenson M. *Sense tagging: semantic tagging with a lexicon* Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?, Washington, D.C. (1997).