Amélioration de liens entre acceptions par fonctions lexicales vectorielles symétriques

Didier Schwab, Mathieu Lafourcade et Violaine Prince
LIRMM
Laboratoire d'informatique, de Robotique
et de Microélectronique de Montpellier
MONTPELLIER - FRANCE.
{schwab,lafourca,prince}@lirmm.fr
http://www.lirmm.fr/ ~{schwab, lafourca, prince}

Mots-clefs – Keywords

représentation thématique, vecteurs conceptuels, fonctions lexicales, acceptions
thematic representation, conceptuals vectors, lexical functions, acceptions

Résumé - Abstract

Dans le cadre du projet Papillon qui vise à la construction de bases lexicales multilingues par acceptions, nous avons défini des stratégies pour peupler un dictionnaire pivot de liens interlingues à partir d'une base vectorielle monolingue. Il peut y avoir un nombre important de sens par entrée et donc l'identification des acceptions correspondantes peut être erronée. Nous améliorons l'intégrité de la base d'acception grâce à des agents experts dans les fonctions lexicales comme la synonymie, l'antonymie, l'hypéronymie ou l'holonymie. Ces agents sont capable de calculer la pertinence d'une relation sémantique entre deux acceptions par les diverses informations lexicales récoltées et les vecteurs conceptuels. Si une certaine pertinence est au-dessus d'un seuil, ils créent un lien sémantique qui peut être utilisé par d'autres agents chargés par exemple de la désambiguïsation ou du transfert lexical. Les agents vérifiant l'intégrité de la base cherchent les incohérences de la base et en avertissent les lexicographes le cas échéant.

In the framework of the Papillon project, we have defined strategies for populating a pivot dictionnary of interlingual links from monolingual vectorial bases. There are quite a number of acceptions per entry thus, the proper identification may be quite troublesome and some added clues beside acception links may be useful. We improve the integrity of the acception base through well known semantic relations like synonymy, antonymy, hyperonymy and holonymy relying on lexical functions agents. These semantic relation agents can compute the pertinence of a semantic relation between two acceptions thanks to various lexical informations and conceptual vectors. When a given pertinence score is above a threshold they create a semantic link which can be walked through by other agents in charge of WSD ot lexical transfert. Base integrity agents walk throw the acceptions, look for incoherences in the base and emit warning toward lexicographs when needed.

1 Introduction

La recherche en représentation de sens est un important problème qui a été abordé selon plusieurs approches. Notre équipe travaille actuellement sur l'analyse thématique de textes et la désambiguïsation lexicale (Lafourcade, 2001). Nous construisons un système capable d'apprentissage automatique basé sur les vecteurs conceptuels. Les vecteurs contiennent les idées associées aux mots ou expressions. Le système d'apprentissage construit ou révise automatiquement les vecteurs conceptuels à partir de définitions en langage naturel contenues dans les dictionnaires à usage humain. Dans le cadre du projet Papillon, nous avons défini des stratégies pour peupler un dictionnaire pivot de liens interlingues (des acceptions) à partir d'une base vectorielle monolingue. L'architecture générale a été décrite dans (Sérasset and Mangeot, 2001) et (Mangeot, 2001). Un dictionnaire pivot (que nous nommerons aussi dictionnaire par acceptions) peut être utilisé avantageusement pour la désambiguïsation et le transfert lexical. Il y a un certain nombre de sens par entrée (environ 5 sens dans nos expériences pour le français) et donc l'identification des acceptions correspondantes peut être erronée. L'amélioration de l'intégrité de la base d'acceptions peut être réalisée grâce à des "agents experts" en relations sémantiques comme la synonymie, l'antonymie, l'hypéronymie ou l'holonymie. Ces agents peuvent calculer la pertinence d'une relation sémantique entre deux acceptions en conjuguant diverses informations lexicales et les vecteurs conceptuels. Lorsqu'un taux de pertinence est au-dessus d'un certain seuil, ils peuvent matérialiser un lien sémantique qui peut être utilisé par d'autres agents en charge de la désambiguïsation ou du transfert lexical. Les agents vérifiant l'intégrité de la base cherchent les incohérences de la base et en avertissent les lexicographes le cas échéant. Dans cet article, nous présentons, dans un premier temps, le modèle des vecteurs conceptuels puis celui du dictionnaire par acception. Ensuite nous présenterons les diverses fonctions lexicales et nous montrerons comment les agents peuvent les utiliser pour créer des liens ou les évaluer.

2 Vecteurs conceptuels

Nous représentons les aspects thématiques des segments textuels (documents, paragraphes, syntagmes, etc) par des vecteurs conceptuels. Les vecteurs ont été utilisés en informatique documentaire pour la recherche d'information (Salton et MacGill, 1983). Leur emploi pour la représentation du sens est plus le fait du modèle LSI (Deerwester et al, 90) issue de l'analyse sémantique latente en psycho-linguistique. En informatique, et de façon presque concurrente, c'est à partir de (Chauché, 90) 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 dont nous faisons l'hypothèse, les concepts, il est possible de construire des vecteurs (dits conceptuels) et de les associer à des items lexicaux[1]. 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 attachées des interprétations linguistiques raisonnables. Les concepts sont donnés a priori. Dans notre expérimentation sur le français nous utilisons (Larousse, 1992) dans lequel sont définis 873 concepts. L'hypothèse principale du thésaurus, que nous adoptons ici, est que cet ensemble constitue un espace générateur pour les termes et leurs sens. D'une façon plus générale, n'importe quel sens peut s'y projeter selon le principe suivant.

[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, 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*).

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 de ‹*vie*› peuvent être projetés sur les concepts suivants (les *CONCEPT*⌈*intensité*⌋ sont ordonnés par valeurs décroissantes) : V‹*vie*› = (*VIE*⌈*0.7*⌋, *NAISSANCE*⌈*0.48*⌋, *ENFANCE*⌈*0.46*⌋, *MORT*⌈*0.43*⌋, *VIEILLESSE*⌈*0.41*⌋, ...) En pratique, plus $\mathcal{C}$ est large, 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 préfèrerons sélectionner les termes thématiquements proches, le *voisinage* (noté $\mathcal{V}$). Par exemple, pour ‹*vie*› : $\mathcal{V}$(‹*vie*›) : ‹*vie*›, ‹*vivant*›, ‹*en vie*›, ‹*naître*›, ... Cette opération est réalisée à l'aide de la distance angulaire.

2.1 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 "$\cdot$" désignant le produit scalaire. Nous supposons ici que les composants des vecteurs sont positifs ou nuls, 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 $D_A(X, Y) \geq \frac{\pi}{4}$, 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 vraie 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(‹*automotrice*›))=0.22 ($13°$)
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(‹*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 synonymes de ‹*locomotive*›, ce qui explique les deux résultats suivants. Le peu de rapports entre ‹*locomotive*› et ‹*rhododendron*› explique l'écart entre leur vecteurs. Dans le dernier exemple, l'angle peu important entre ‹*locomotive*› et ‹*guépard*› au regard de celui entre ‹*locomotive*› et ‹*rhododendron*› se comprend si on se rappelle que D_A est une distance thématique et non une distance ontologique. 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. Seule une expertise humaine est capable de juger de la pertinence des vecteurs (si les résultats renvoyés sont cohérents avec la langue).

2.2 Opérations sur les vecteurs

Les opérations suivantes ont étés définies dans (Lafourcade et Prince, 2001), nous les rappelons brièvement.

$\oplus$ est la somme vectorielle normalisée définie par $V = X \oplus Y \quad | \quad v_i = \frac{x_i + y_i}{\|V\|}$
$\otimes$ est le produit terme à terme normalisé défini par $V = X \otimes Y \quad | \quad v_i = \sqrt{x_i \times y_i}$
Γ est la contextualisation faible définie par $\Gamma(X, Y) = X \oplus (X \otimes Y)$

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

2.3 Construction des vecteurs conceptuels

La construction des vecteurs conceptuels se fait à partir de définitions extraites de diverses sources (dictionnaires, listes de synonymes, indexations manuelles, ...). Cette méthode d'analyse construit, à partir de vecteurs conceptuels déjà existants et de nouvelles définitions, de nouveaux vecteurs. Il est nécessaire d'effectuer l'amorçage du système d'apprentissage à partir d'un noyau constitué de vecteurs calculés au préalable pour les termes les plus courants. Les items lexicaux de ce noyau sont considérés comme pertinents. Cet ensemble constitue la base d'items lexicaux à partir de laquelle a démarré l'apprentissage. Nous cherchons à mettre au point un apprentissage qui soit le plus cohérent possible afin d'obtenir une base augmentée pertinente. Une des manières d'améliorer cette cohérence est de tirer parti des relations sémantiques[3] qui existent entre les items. Une autre manière consiste à faire de l'apprentissage sur des langues différentes puis de comparer les vecteurs grâce à des dictionnaires interlingues. D'autres moyens pour améliorer les vecteurs existent ou peuvent être envisagés.

2.4 Agents

Dans cet article, nous considérons comme agent *toute entité humaine ou artificielle qui peut agir sur la base d'acception.* L'action peut se faire par la création ou la suppression de liens entre acceptions, la création ou la suppression d'acceptions.

Un *agent spécialiste d'un domaine* (appelé aussi *agent expert*) est un agent qui a une compétence particulière dans un certain domaine (relation sémantique, analyse sémantique, désambiguïsation, ...). Les agents spécialistes d'un autre domaine sont appelés *agents non-experts* ou *agents non-spécialistes*. Par exemple, un agent spécialiste de l'antonymie est considéré comme non-spécialiste de tous les autres domaines (en particulier des autres relations sémantiques).

3 Description de la base lexicale

3.1 Acceptions

Une acception est un sens particulier d'un mot, admis et reconnu par l'usage. Il s'agit d'une unité sémantique propre à une langue donnée (Sérasset and Mangeot, 2001). Par exemple, l'item lexical ‹*botte*› a au moins trois acceptions, la *chaussure*, *amas de paille* ou le *coup*. Contrairement aux items lexicaux, les acceptions sont donc monosémiques. Cela a des conséquences sur les relations sémantiques, elles n'ont plus besoin d'être appréciées en contexte. Dans ce cas de figure, la synonymie est une relation d'équivalence et les relations hiérarchiques sont transitives.

3.2 Base d'acceptions

Le modèle est constitué de deux parties : des dictionnaires monolingues et une base intermédiaire, celle des acceptions. Les entrées de chaque dictionnaire monolingue sont reliées aux acceptions correspondantes. Ces acceptions font donc office de pivots entre les items de chaque langue (cf fig. 1 droite)

[3] hypéronymie/hyponymie et méronymie/holonymie qui sont de type hiérarchique et synonymie et antonymie qui sont de type symétrique.

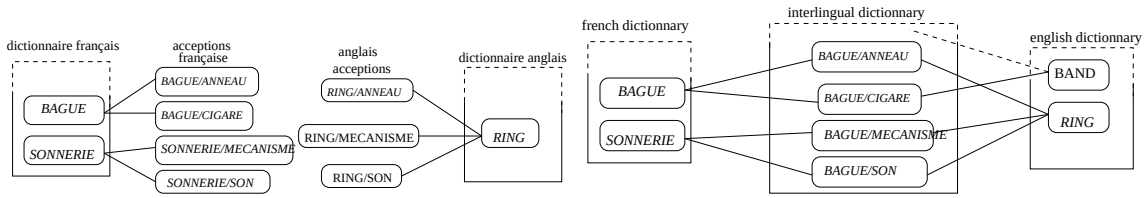

Figure 1: Architecture de la base lexicale. À gauche, les dictionnaires monolingues et leurs acceptions. À droite, la base d'acceptions interlingues où les acceptions "identiques" de chaque langue ont étés fusionnées.

Dans chaque base vectorielle monolingue, chaque item est relié à une acception de sa langue (cf fig. 1 gauche). Chacune des acceptions a son propre vecteur conceptuel. Les acceptions interlingues sont construites à partir de ces acceptions monolingues (Lafourcade, 2002). La base d'acceptions est de grande taille (plus de 500000 acceptions) et est construite par différents agents qui peuvent être humains ou artificiels. Nous considérons qu'il est très difficile de maintenir intacte l'intégrité de la base sans contrôle : un agent (certainement humain) peut toujours créer une nouvelle acception même si une acception adéquate existe déja. Un moyen de vérifier cette intégrité est de chercher les liens sémantiques entre acceptions. Nous considérons plusieurs liens sémantiques dans le lexique et nous montrons comment les utiliser pour vérifier l'intégrité de la base ou comment un agent non-spécialiste peut utiliser les liens pour en évaluer d'autres.

4 Liens sémantiques

Les relations sémantiques entre items lexicaux structurent le lexique sur le plan paradigmatique. Ces relations sont de deux types : les relations hiérarchiques (hyponymie/hypéronymie, méronymie/holonymie) et les relations d'équivalence/opposition (synonymie, antonymie). Elles sont souvent décrites comme des relations booléennes, i.e, elles existent entre deux items ou non (Polguère, 2001).

Dans le cas des acceptions, monosémiques par définitions, les relations hiérarchiques sont transitives et la synonymie peut être considérée comme une équivalence. Si nous matérialisons toutes les relations, nous nous heurtons au problème de l'explosion des liens. Il s'agit, non seulement, d'un problème matériel (taille en mémoire) mais aussi d'un problème qui met en avant la question de l'interêt des liens : leur pouvoir de discrimination est inversement proportionnel à leur nombre. Par exemple, toutes les acceptions de noms seraient liées à un terme général comme ‹*objet concret*› pour un lien hypéronymique. Par exemple, pour ‹*chat*›, ‹*félin*› semble être un meilleur hypéronyme que ‹*animal*› ou ‹*mammifère*›. Pour éviter ce problème, et afin de pouvoir comparer deux liens, nous utilisons des *relations sémantiques valuées*.

4.1 Relations sémantiques valuées

Les relations sémantiques valuées (RSV) ne sont pas booléennes et ont une valeur qui exprime la pertinence d'une relation entre deux items lexicaux. Une RSV $\mathcal{R}$ est une relation qui donne, pour deux items, une valeur entre 0 et 1 : $\mathcal{R} : w^2 \rightarrow [0, 1]$ où w est l'ensemble des items lexicaux. Plus la valeur est proche de 1, plus la relation entre les deux items est pertinente, plus la valeur est proche de 0, moins la relation entre les deux items est pertinente. Si la valeur est de 0, nous pouvons considérer que la relation ne s'applique pas entre les deux termes. La valeur de la relation peut être vue comme la probabilité que la relation existe. Les définitions des différentes relations sont données dans les paragraphes correspondants.

4.2 Création et suppression de liens

Nous voulons ajouter des liens sémantiques à la base d'acceptions afin, non seulement de constituer un réseau sémantique mais également pour pouvoir vérifier son intégrité. Les agents capables d'évaluer une relation sémantique valuée peuvent créer des liens sémantiques valués si la valuation est supérieure à un seuil s. Par exemple, si un agent expert d'antonymie évalue l'antonymie entre ‹*froid*› et ‹*chaud*› à une valeur n supérieure à s, il construit un lien sémantique entre les deux acceptions valué à n. Le seuil n'est pas fixé en avance et il évolue constament en fonction du nombre de liens déjà construits. Le système apprend en fonction des nouvelles données (nouveaux dictionnaires monolingues ou bilingues) ou en révisant les anciennes données et donc les agents doivent régulièrement recalculer les relations et si la condition pour préserver le lien (être supérieur à s) n'est plus respectée alors le lien est détruit. Ces liens matérialisés peuvent être utilisés par des agents qui ne peuvent pas calculer ces relations sémantiques par eux même mais qui sont capables d'évaluer rapidement les liens à partir de simples règles. Par exemple, un agent peut évaluer les propriétés de transitivité d'un hypéronyme pour évaluer la valeur de $Hyp(A, C)$ à partir $Hyp(A, B)$ et de $Hyp(B, C)$. Cela peut être utile si la base d'acceptions est utilisée pour une tâche de désambiguïsation par exemple. En aucun cas, un agent non-expert peut construire un lien.

Bien qu'un certain nombre de pistes concernant les relations hiérarchiques aient été lancées dans les premières pages de cet article, nous ne nous intéresserons dans la suite qu'aux relations sémantiques symétriques, antonymie et synonymie, ainsi qu'aux règles qui peuvent être appliquées pour vérifier l'intégrité de la base et déduire de nouvelles relations à partir de celles déjà existantes.

4.3 Synonymie

La synonymie est la relation sémantique qu'il existe entre deux items lexicaux qui peuvent, dans un contexte donné, exprimer le même sens. Par exemple, ‹*avion*› et ‹*aéroplane*› sont synonymes. Contrairement aux unités lexicales, les acceptions sont monosémiques par définition. Dans ce contexte, nous pouvons définir la synonymie comme *la relation sémantique qui existe entre deux acceptions qui expriment le même sens*.

4.3.1 Synonymie relative

Dans (Lafourcade et Prince, 2001), la synonymie est étudiée à travers la notion de synonymie relative. Nous définissons la fonction de synonymie relative Syn_R entre trois vecteurs A, B, C, ce dernier étant le référent, comme suit :

$$Syn_R(A, B, C) = D_A(\Gamma(A, C), \Gamma(B, C)) = D_A(A \oplus (A \otimes C), B \oplus (B \otimes C))$$

On cherche à tester la proximité thématique de deux sens (A et B), chacun augmenté de ce qu'il a de commun avec un troisième (C). La synonymie relative est une distance, elle vérifie donc la réflexivité, la symétrie et la transitivité.

4.3.2 Relation sémantique valuée de la synonymie

Soit la relation sémantique valuée de la synonymie définie par $Syn(X, Y) = 1 - \frac{2}{\pi} Syn_R(X, Y, X \oplus Y)$. Il s'agit du passage de l'intervalle $[0, \frac{\pi}{2}]$ à l'intervalle $[0, 1]$ de la synonymie relative. Cette transformation inverse le domaine image de façon linéaire. Pour une transformation vers $[0, 1]$, nous aurions pu

utiliser le cosinus mais on se garde bien de le faire car on souhaite focaliser le pouvoir de discrimination dans les faibles valeurs de l'angle.

4.3.3 Calcul de la synonymie par des agents non-spécialistes

Un agent non-expert peut évaluer la synonymie entre deux acceptions grâce à la formule suivante :

$$Syn(A, C) = Min_{i \in I}(Min(Syn(A, X_i), Syn(X_i, B)))$$

avec I, l'ensemble des items reliés à la fois à A et B.

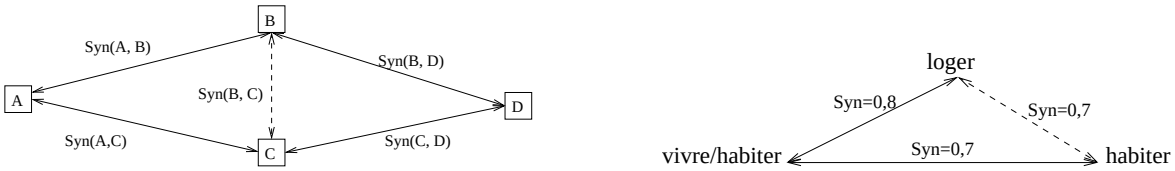

Figure 2: Evaluation de la relation de synonymie.

S'il existe un chemin allant de A à B, nous considérons que le RSV entre A et B est le plus petit RSV du chemin. Lorsque plusieurs chemins sont possibles, la même idée que pour les relations hiérarchiques est adoptée, nous choisissons le plus mauvais chemin (le moins probable) pour évaluer le RSV. Dans la partie droite de la figure 2, la synonymie entre ‹*habiter*› et ‹*loger*› est évaluée grâce à la synonymie entre ‹*vivre/habiter*› et ‹*loger*› et la synonymie entre ‹*loger*› et ‹*habiter*›.

4.4 Antonymie

(Schwab et all., 2002) propose une définition de l'antonymie compatible avec le modèle vectoriel utilisé. Le transfert aux acceptions ne modifie pas cette définition. *Deux acceptions sont en relation d'antonymie si on peut exhiber une symétrie de leurs traits sémantiques par rapport à un axe*. Nous considérons que les relations d'antonymie dépendent du type de support de symétrie. Pour une acception, il peut exister plusieurs types de symétrie possibles comme il peut ne pas y en avoir d'évident si le support de symétrie ne peut être trouvé. Plusieurs sortes de support peuvent être distingués : (i) une propriété affectant une valeur étalonnable (*chaud* et *froid* qui sont des valeurs symétriques de température) (ii) l'application d'une propriété (applicable/inapplicable, présence/absence), l'existence d'une propriété ou d'un élément considéré comme symétrique par l'usage (e.g. *soleil/lune*), ou par des propriétés naturelles ou physiques des objets considérés (e.g. mâle /femelle, tête/pied, ...).

4.4.1 Acceptions sans antonymes

Une conséquence importante de notre définition est que tout vecteur conceptuel peut avoir un vecteur antonyme. En effet, pour un axe donné, tout vecteur a un symétrique. La linguistique classique considère que certains termes, n'ont pas d'antonymes avérés[4]. Les idées principales constituant ces acceptions, autrement dit, les concepts, ne sont pas obligatoirement opposables. Dans un espace géométrique, un point qui n'a pas d'autre symétrique se trouve sur l'axe de symétrie. De même, dans notre formalisme, les idées qui ne sont pas opposables sont sur l'axe de symétrie et donc l'antonyme d'un item lexical qui ne possède pas d'antonyme avéré est l'item lexical lui-même. Nous appelons cela, *la propriété des points fixes.* En pratique, les concepts possèdent plus facilement cette propriété

[4]C'est le cas, par exemple, des objets matériels comme ‹*table*›, ‹*voiture*› ou ‹*porte*›.

que les acceptions. Ces derniers, en se projetant sur plusieurs concepts (l'immense majorité des mots de la langue étant polysème), peuvent, dans certains contextes, hériter de la capacité d'opposition de certains concepts. Ainsi par exemple, en antonymie scalaire, une ‹*Ferrari*›, bien que sorte d'*AUTOMOBILE*, concept sans antonyme, se projette aussi sur une notion de *RAPIDITÉ* qui, elle, est opposable. C'est pourquoi on peut très bien imaginer une ‹*deux chevaux*› comme possible antonyme d'une ‹*Ferrari*›.

À partir de ces spécifications, nous avions défini dans (Schwab et all., 2002) une fonction $Anti_R$ simulant l'antonymie. Nous utilisons cette fonction pour créer une mesure d'évaluation de l'antonymie.

4.4.2 Mesure d'évaluation de l'antonymie

Cette mesure permet de vérifier si deux items lexicaux (ou acceptions) peuvent être antonymes. Soit A et B deux vecteurs, la question est précisément de savoir s'ils peuvent raisonnablement être antonymes dans un contexte C. La mesure d'antonymie $Manti_{Eval}$ est la mesure en radians de l'angle entre la somme de A et B et la somme de $Anti_{c_R}(A, C)$ et $Anti_{c_R}(B, C)$. Ainsi, nous avons :

$$Manti_{Eval} = D_A(A \oplus B, Anti_R(A, C) \oplus Anti_R(B, C))$$

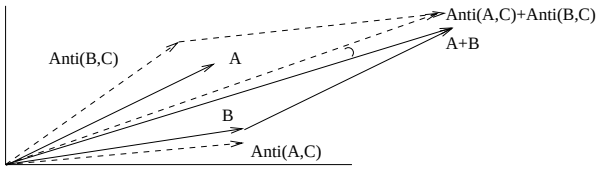

Figure 3: représention géométrique en 2 dimensions de la mesure d'évaluation de l'antonymie $Manti_{Eval}$

La mesure d'antonymie est une pseudo-distance. Elle vérifie les propriétés de réflexivité, symétrie et inégalité triangulaire uniquement dans le sous ensemble des items qui n'ont pas d'antonymes. Dans le cas général, elle ne vérifie pas la réflexivité. Les composantes des vecteurs conceptuels sont positives et nous avons la propriété $Dist_{anti} \in [0, \frac{\pi}{2}]$. Plus la mesure est petite, plus les deux items lexicaux sont antonymes dans le contexte. En revanche, ce serait une erreur de considérer que deux antonymes seraient à une distance avoisinant $\pi/2$. Deux items lexicaux à $Manti_{Eval} = \pi/2$ l'un de l'autre n'ont aucune idée en commun[5], ce qui n'est pas le cas de deux antonymes qui on en commun les idées non opposables ou celles qui le sont mais dont l'activation est proche. Ils ne s'opposent que par certaines activations de concepts. Une distance de $\pi/2$ entre deux items lexicaux devrait être plutôt interprété comme une sorte d'anti-synonymie. Ce résultat confirme le fait que l'antonymie n'est pas exactement l'inverse de la synonymie mais lui est très liée. L'antonyme d'un item ‹*m*› n'est pas un mot qui ne partage aucune idée avec ‹*m*› mais un item qui s'oppose à ‹*m*› sur certaines idées.

4.4.3 Relation sémantique valuée de l'antonymie

Nous définissons la relation sémantique valuée de l'antonymie par $Anti(X, Y) = 1 - \frac{2}{\pi} Manti_{Eval}(X, Y)$ Il s'agit de la conversion linéaire de l'intervalle $[0, \frac{\pi}{2}]$ à l'intervalle $[0, 1]$.

[5] ce cas de figure est purement théorique, il n'existe dans aucune langue deux items lexicaux qui ne partagent aucune idée.

5 Amélioration de l'intégrité de la base d'acceptions

5.1 Schéma général de cohérence

Le schéma général de cohérence doit être vérifié dans la base d'acceptions (fig. 4) sinon la base n'est pas cohérente. Si A est antonyme de B et B antonyme de C alors A et B sont synonymes. De même,

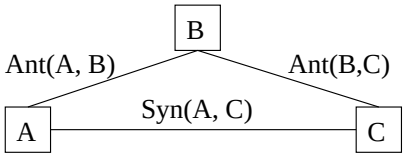

Figure 4: Schéma général de cohérence

si A et B sont antonymes et A et C synonymes alors B et C sont antonymes. Des agents spécialistes cherchent ces schémas et avertissent les lexicographes en cas d'incohérences. Les lexicographes peuvent alors indiquer si une acception doit être divisée ou si un lien ne devrait pas être matérialisé.

5.2 Évaluation de la synonymie et de l'antonymie

Le schéma général de cohérence (fig. 4) peut aussi aider les agents non-spécialistes pour évaluer un lien non-matériel. Si A est antonyme de B et B antonyme de C alors nous avons A et B synonymes. Dans le cas général, nous avons :

$$Syn(A, C) = Min_i(Min(Ant(A, X_i), Ant(X_i, C)))$$

La figure 5 montre un exemple d'évaluation de la synonymie à partir de l'antonymie.

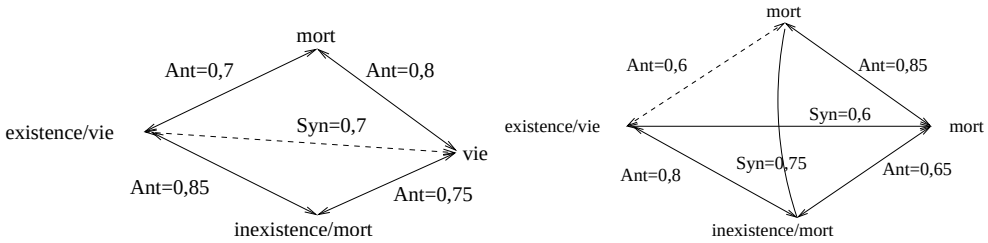

Figure 5: Exemple d'évaluation de la synonymie (à gauche) et de l'antonymie (à droite)

De la même manière, nous pouvons aussi évaluer l'antonymie:

$$Ant_i(A, C) = Min_i(Min(Syn(A, X_i), Ant(X_i, C)))$$

La figure 5 montre un exemple d'évaluation de la synonymie à partir de l'antonymie.

6 Conclusion et perspectives

Dans cet article, nous avons présenté une manière d'améliorer l'intégrité d'une base d'acceptions grâce aux relations sémantiques symétriques bien connues que sont la synonymie et l'antonymie. Nous avons présenté les relations sémantiques valuées (RSV) qui peuvent être comparées à la probabilité que la relation existe entre deux items ou acceptions. Les RSV sont calculées par des agents spécialistes grâce à diverses informations lexicales et aux vecteurs conceptuels pour créer des liens

matérialisés entre deux acceptions si ce RSV dépasse un certain seuil. Nous avons montré comment des agents non-spécialistes (en charge du transfert lexical ou de la désambiguïsation du sens) peuvent évaluer des liens non-matérialisés à partir de liens matérialisés. Les agents vérifiant l'intégrité de la base utilisent ces liens pour chercher les incohérences de la base et avertissent les lexicographes le cas échéant. Il nous reste encore à prolonger ces travaux en cherchant à automatiser ces corrections d'incohérences. Ces travaux doivent aussi être approfondis en ce qui concerne les relations hiérarchiques comme l'holonymie ou l'hypéronymie. Notre équipe travaille actuellement sur le calcul de RSV concernant ces relations ainsi que sur les moyens de les utiliser pour améliorer la cohérence de bases d'acceptions.

Références

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

Deerwester S. et Dumais S., Landauer S., Furnas G., Harshman R., *Indexing by latent semantic anlysis.* In Journal of the American Society of Information science, 1990, 416(6), pp 391-407.

Lafourcade M., *Automatically Populating Acception Lexical Database through Bilingual Dictionaries and Conceptual Vectors.* proc. de PAPILLON-2002, Tokyo, Japan, August 2002.

Lafourcade M. et Prince V. *Synonymies et vecteurs conceptuels.* Proc. of Traitement Automatique du Langages Naturel (TALN'2001) (Tours, France, Juillet 2001), pp 233-242.

Lafourcade M . *Lexical sorting and lexical transfer by conceptual vectors.* Proc. of the First International Workshop on MultiMedia Annotation (Tokyo, Janvier 2001), 6 p.

Larousse. *Le Petit Larousse Illustré 2001.* Larousse, 2001.

Larousse. *Thésaurus Larousse - des idées aux mots, des mots aux idées.* Larousse, ISBN 2-03-320-148-1, 1992.

Lyons J. *Semantics.* Cambridge : Cambridge University Press, 1977.

. Mangeot-Lerebours M. *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingues.* thèse de doctorat de l'Université Joseph Fourier, 2001.

Morin, E. *Extraction de liens sémantiques entre termes à partir de corpus techniques.* Thèse de doctorat de l'Université de Nantes, 1999.

Muehleisen V.L. *Antonymy and semantic range in english.* Northwestern university Phd, 1997.

Palmer, F.R. *Semantics : a new introduction.* Cambridge University Press, 1976.

Polguére A. *Notions de base en lexicologie.* Observatoire de linguistique sens-texte, 2001.

Thesaurus of English Words and Phrases. Longman, London, 1852.

Salton G. et MacGill M.J. *Introduction to modern Information Retrieval.* McGraw-Hill, New-York, 1983.

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

Sérasset G., Mangeot M. *Papillon lexical databases project: monolingual dictionnaries & interlingual links.* NLPRS 2001 processings, pp 119-125.