Amélioration de la représentation sémantique lexicale par les vecteurs conceptuels: le rôle de l'antonymie

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}

Abstract

In the framework of research in meaning representations in NLP, we focus our attention on thematic aspects and conceptual vectors. The learning strategy associated with conceptual vectors relies on the morphosyntaxic analysis of human usage dictionary definitions linked to vector propagation. This analysis currently doesn't take into account negation phenomena. This work aims at studying the antonymy aspects of negation, in the larger goal of its integration into the thematic analysis. After a linguistic presentation of the antonymy, we present a model based on the idea of symmetry compatible with conceptual vectors. Then, we define antonymy functions which allow the construction of an antonymous vector and the enumeration of potantially antinomic lexical items. Finally, we introduce some measure functions, which evaluate how a given vector might accept an antonym and how a given word is an acceptable antonym of another term.

Résumé

Dans le cadre de recherches sur le sens en traitement automatique du langage, nous nous concentrons sur la représentation de l'aspect thématique des segments textuels sous la forme de vecteurs conceptuels. Le système d'apprentissage des vecteurs conceptuels est fondé sur l'analyse de définitions issues de dictionnaires à usage humain (Schwab, 2001) Cette analyse ne gère pas les phénomènes liés à la négation. Cet article vise à étudier l'antonymie comme expression de l'opposition d'idées et montre comment l'utiliser dans le raffinement du processus de construction des vecteurs. Nous abordons d'abord l'antonymie du point de vue linguistique et introduisons un modèle basé sur une idée de symétrie compatible avec les vecteurs conceptuels. Nous définissons ensuite des fonctions d'antonymie qui permettent d'obtenir le vecteur antonyme ou une énumération des items potentiellements antonymes. Nous présentons des mesures permettant de savoir si un vecteur peut raisonnablement posséder un antonyme ou si un mot est l'antonyme d'un autre.

Mots-clés : thematic representation, conceptuals vectors, complementary, scalar and dual antonymy, potential antonymy mesure, antonymy evaluation mesure

1. Introduction

Dans le cadre de la représentation du sens en TALN, nous nous intéressons plus particulièrement aux aspects thématiques et aux vecteurs conceptuels (Lafourcade, 2001). Certaines formes syntaxiques sont indirectement porteuses de sens. Elles peuvent être modélisées grâce à la théorie sens-texte et aux fonctions lexicales (Mel'čuk et al., 1995). Nous nous focalisons sur l'antonymie comme un aspect possible de la négation en mettant l'accent sur son côté opposition d'idées. L'équipe de TAL du LIRMM étudie actuellement un système d'analyse de textes dans leurs aspects thématiques et de désambiguïsation lexicale basée sur les vecteurs conceptuels. Ces vecteurs représentent les idées associées aux mots, aux expressions et de façon générale,

à tout segment textuel. Le système d'apprentissage des vecteurs les construit ou les révise à partir de définitions en langage naturel extraites de dictionnaires à usage humain. Cette stratégie permet de s'assurer de données lexicales linguistiquement maîtrisés (Véronis and Ide, 1992). Cependant, les définitions "négatives" restent problématique. Dans (Larousse, 2001), par exemple, ‹*inexistant*› est cherché à définir par *"qui n'existe pas"*. Jusqu'à présent, notre analyseur sémantique automatique des définitions ne gère pas bien ce type de définitions. Il n'analyse pas bien non plus les définitions qui contiennent des antonymes comme ‹*action*›: "antonymes: ‹*réaction*›, ‹*inaction*›" Nous avons donc défini une fonction d'antonymie pour les vecteurs. Ainsi, ‹*inexistant*› n'est plus directement calculé à partir du vecteur ‹*existant*› mais à partir du vecteur opposé à ‹*existant*›. Cela a permis d'améliorer le système d'apprentissage et par conséquent, la précision de tous les vecteurs mais également, l'augmentation de la qualité du processus global d'analyse de textes. De manière plus générale, une fonction d'antonymie peut aussi aider à trouver une thématique opposée qui peut être utilisée dans toute application générative de textes (recherche des idées opposées, paraphrase, résumé, etc).

2. Définitions et caractérisation de l'antonymie

2.1. Généralités

La définition de l'antonymie fournie par (Larousse, 1991) est la suivante : *Un antonyme est un mot qui a un sens opposé à celui d'un autre. Les antonymes, ou contraires, sont des mots appartenant obligatoirement à la même classe grammaticale ("grand" est l'antonyme de "petit" et non celui de "petitesse") et s'opposant par un ou plusieurs traits sémantiques, les autres étant communs. Par exemple, "monter" et "descendre" possèdent en commun le trait "déplacement vertical" et s'opposent par les traits "vers le haut" et "vers le bas". L'antonymie peut donc se définir comme une relation d'incompatibilité entre deux termes. Elle est, à cet égard, l'exact opposé de la synonymie.* Cette définition de l'antonymie doit être revue à la lumière de la modélisation vectorielle que nous proposons. En effet, si on veut caractériser la construction de l'antonyme d'un concept, il est préférable d'utiliser la notion de *symétrie* plutôt que celle d'incompatibilité. La symétrie se décline alors de différentes manières, selon la nature de son support. On distingue, comme supports :

- une **propriété** affectant une valeur étalonnable (valeur élevée, valeur faible) : par exemple, *chaud, froid* sont des valeurs symétriques de température;
- l'**application d'une propriété** (applicable/non applicable, présence/absence) : par exemple, *informe* est antonyme de tout ce qui a une forme, *insipide, incolore, inodore, etc.* de tout ce qui pourrait avoir saveur, couleur, odeur. . .
- 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, . . .*).

Notre idée est que les constructions d'antonymes sont dépendantes du type de support de symétrie. Il peut alors exister plusieurs types d'antonymes pour un même terme, comme il peut ne pas en exister d'évidents, si la symétrie n'est pas immédiatement décelable. En tant que fonction lexicale, comparée à la synonymie, on peut dire que si la synonymie est la recherche de la ressemblance avec comme test la substitution (*x est synonyme de y si x peut "remplacer" y*), l'antonymie est la recherche de la symétrie avec comme test la recherche du support de la symétrie (*x est antonyme de y s'il existe un support de symétrie t tel que x symétrique de y par*

rapport à t). Par exemple, ‹*chaud*› est antonyme de ‹*froid*› car ‹*température*› offre un support de symétrie.

De même que pour la synonymie (Lafourcade and Prince, 2001), il n'existe pas d'antonymes absolus, i.e. deux mots qui seraient antonymes l'un de l'autre quel que soit le contexte. L'antonymie s'apprécie toujours en contexte. Par exemple, *frais* peut être le contraire de *tiède, chaud, racorni, flétri, maladif, rassis, confit, sec, surgelé, pourri, ...*

Bien qu'il existe plusieurs types d'antonymie, nous n'en considèrerons qu'une seule dans ce qui suit, l'antonymie complémentaire. Nous avons défini d'autres types d'antonymies (Schwab, 2001). La méthode est la même quel que soit le type, elles ne diffèrent qu'au niveau des listes d'antonymes (5.3).

3. L'antonymie complémentaire

Les antonymes complémentaires sont les couples comme *pair/impair, présence/absence*.

Ce nombre est pair ⇒ Ce nombre n'est pas impair. Ce nombre n'est pas pair ⇒ Ce nombre est impair
Ce nombre est impair ⇒ Ce nombre n'est pas pair Ce nombre n'est pas impair ⇒ Ce nombre est pair

En terme de logique,

$$\forall x \quad P(x) \Rightarrow \neg Q(x) \qquad \forall x \quad \neg P(x) \Rightarrow Q(x)$$
$$\forall x \quad Q(x) \Rightarrow \neg P(x) \qquad \forall x \quad \neg Q(x) \Rightarrow P(x)$$

L'affirmation de l'un des termes correspond à la négation de l'autre, il s'agit d'une relation de disjonction exclusive. Cette antonymie présente deux types de symétrie, une **symétrie de valeur dans un système booléen** comme dans l'exemple précédent et une **symétrie quant à l'application d'une propriété** (le *noir* est le manque de couleur, il est donc "opposé " à toute autre couleur ou combinaison de couleur).

4. Vecteurs conceptuels

Dans le cadre de la recherche sur la sémantique en TALN, nous représentons les aspects thématiques des segments textuels (documents, paragraphes, syntagmes, etc) par des vecteurs conceptuels. Ce modèle inspiré de (Salton and McGill, 1983) et (Deerwester et al., 1990) est basé sur la projection dans le formalisme mathématique d'espace vectoriel de la notion linguistique de champ sémantique. À partir d'un ensemble de notions élémentaires dont nous faisons l'hypothèse, les concepts, il est possible de construire des vecteurs (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 définis selon un thésaurus (dans notre expérimentation sur le français nous utilisons (Larousse, 1992) dans lequel sont défini 873 concepts). L'hypothèse principale est que cet ensemble constitue un espace générateur non libre pour les termes et leurs sens. N'importe quel mot ou, de façon plus générale, sens peut s'y projeter selon le principe ci-après.

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*).

4.1. Principe

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 ‹*porte*› peuvent être projetés sur les concepts suivants (les *CONCEPT*⌈*intensité*⌋ sont ordonnés par valeurs décroissantes):

V‹*porte*› = (*OUVERTURE*⌈*0.36*⌋, *BARRIÈRE*⌈*0.35*⌋, *FERMETURE*⌈*0.33*⌋, *LIMITE*⌈*0.32*⌋, *EXTÉRIEUR*⌈*0.31*⌋, *CONTENANT*⌈*0.3*⌋ (*INTÉRIEUR*⌈*0.25*⌋) (*CENTRE*⌈*0.2*⌋) (*REVÊTEMENT*⌈*0.19*⌋) ...

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[2], l'énumération des concepts activés est longue et il est difficile d'enévaluer la pertinence. En général, nous préfèrerons sélectionner les termes thématiquements proches, le *voisinage* (noté $\mathcal{V}$). Par exemple, les termes les plus proches de ‹*porte*› sont: $\mathcal{V}$(‹*porte*›) : ‹*porte-fenêtre*›, ‹*portail*›, ‹*portière*›, ‹*ouverture*›,... Cette opération est réalisée à l'aide de la distance angulaire.

4.2. 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 généralement que pour une distance $D_A(X, Y) \leq \frac{\pi}{4}$ (45°), X et Y sont thématiquement proche et partage 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°), ils 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, symétrie et inégalité triangulaire. Nous obtenons, par exemple, les angles suivants[3].

D_A(V(‹*mésange*›), V(‹*mésange*›))=0 (0°)	D_A(V(‹*mésange*›), V(‹*train*›))=1.28 (73°)
D_A(V(‹*mésange*›), V(‹*oiseau*›))=0,55 (31°)	D_A(V(‹*mésange*›), V(‹*insecte*›))=0,57 (32°)
D_A(V(‹*mésange*›), V(‹*passereau*›))=0,35 (20°)	D_A(V(‹*mésange*›), V(‹*couleur*›))=0,59 (33°)

Le premier résultat a une interprétation directe, ‹*mésange*› ne peut être plus proche d'autre chose que de lui même. Le fait qu'une mésange soit un oiseau de la famille des passereaux explique les deuxième et troisième résultats. Une mésange n'a que peu de rapport avec un ‹*train*› ce qui explique l'angle plus important. On peut se demander pourquoi l'angle entre ‹*mésange*› et ‹*insecte*› est si peu important. Il faut se rappeler que D_A est une distance thématique et non une distance ontologique. L'examen de la définition de mésange explique le résultat *(Oiseau*

2. Les vecteurs denses sont ceux qui ont peu de coordonnées nulles.

3. Les exemples sont extraits de `http://www.lirmm.fr/~schwab` pour la partie concernant plus spécifiquement l'antonymie et de `http://www.lirmm.fr/~lafourca` pour tous les autres exemples.

passériforme insectivore, long de 10 à 14 cm, au plumage coloré, aux mouvements vifs, dont la plupart des espèces appartiennent au genre Parus). On remarquera que les comparaisons entre les valeurs sont plus significatives que les valeurs elles-mêmes.

Des opérations sur les vecteurs ont été définies, nous allons maintenant présenter celle qui est utilisée dans les fonctions d'antonymie, la somme.

4.3. Somme vectorielle

Soient X et Y deux vecteurs. Leur somme normée est définie par:

$$V = X \oplus Y \quad | \quad v_i = \frac{x_i + y_i}{\|V\|}$$

L'opérateur est idempotent, nous avons $X \oplus X = X$. Le vecteur nul $\vec{\mathbf{0}}$ est un élément neutre de la somme vectorielle. Par définition, nous avons $\vec{\mathbf{0}} \oplus \vec{\mathbf{0}} = \vec{\mathbf{0}}$.. La somme vectorielle est généralisée à un nombre quelconque de vecteurs par:

$$V = \bigoplus_{i=1}^{n} X_i \quad | \quad v_i = \frac{\sum x_i}{\|V\|}$$

4.4. 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 forme, à partir de vecteurs conceptuels déjà existants et de 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. Cet ensemble constitue la base d'items lexicaux sur laquelle se basera l'apprentissage.

5. Fonctions d'antonymie

5.1. Principes et définitions

Nous rappelons que l'objectif de nos travaux est de créer une fonction d'antonymie afin d'améliorer le processus d'apprentissage. Dans ce qui suit, nous allons surtout nous intéresser à la génération d'antonymes qui permet de fournir un bon indice de satisfaction et de ces fonctions. Nous allons donc définir une fonction qui, à partir d'items lexicaux, renverra les N plus proches antonymes à la manière de la fonction de voisinage $\mathcal{V}$.

Nous l'avons vu dans 2, un antonyme s'apprécie toujours en contexte. Pourtant, dans certains cas, ce contexte seul ne nous semble pas suffisant pour déterminer un axe de symétrie à l'antonymie. Prenons l'exemple de l'item lexical ‹*père*›. Dans le contexte ‹*famille*›, il peut être opposable à ‹*mère*› ou ‹*enfant*›. Il peut être pertinent, dans les cas où il ne sert pas d'axe de symétrie, d'affiner le contexte par un vecteur conceptuel qui peut, lui, jouer ce rôle de référent. Dans notre exemple, il faudrait prendre alors comme référent ‹*filiation*›, ‹*mariage*› ou ‹*homme*›. C'est la raison pour laquelle nous avons défini la fonction d'antonymie $AntiLex_S$ qui renvoie les n plus proches antonymes du mot X dans le contexte défini par le mot C en référence au mot R (Schwab, 2001)

$$X, C, R, n \rightarrow AntiLex_S(X, C, R, n)$$

$$X, C, n \rightarrow AntiLex_R(X, C, n) = AntiLex_S(X, C, C, n)$$

$$X, R, n \rightarrow AntiLex_B(X, R, n) = AntiLex_S(X, R, R, n)$$

$$X, n \rightarrow AntiLex_A(X, n) = AntiLex_S(X, X, X, n)$$

Nous avons défini la fonction partielle $AntiLex_R$ car, dans la plupart des cas, le contexte suffit à déterminer un axe de symétrie. $AntiLex_B$ peut être défini si on souhaite déterminer un axe de symétrie et pas un contexte. Cette distinction reste largement théorique car dans la pratique, on a $AntiLex_B = AntiLex_R$. La dernière fonction est la fonction d'antonymie absolue. Son usage est délicat dans le cas de mots polysémiques puisque aucun contexte n'aide à savoir quel sous-sens doit être privilégié sur les autres. Cela augmente la probabilité de ne pas obtenir une réponse satisfaisante.

La fonction $AntiLex_S$ effextue le parcours présenté en figure 1.

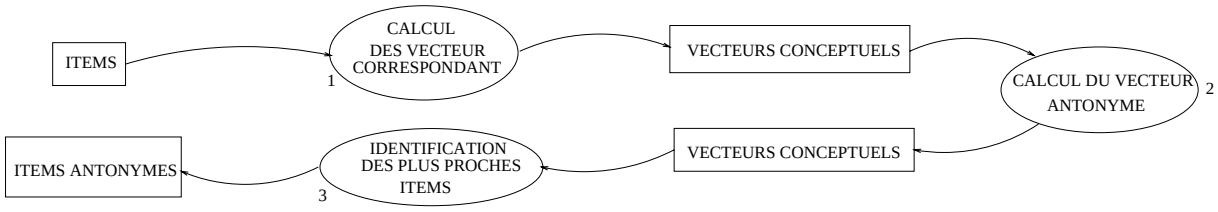

FIG. 1 – *Parcours des fonctions* $AntiLex$

Tout d'abord, nous allons nous intéresser plus particulièrement au calcul du vecteur antonyme, c'est à dire à la fonction d'antonymie au niveau des vecteurs conceptuels (étape 2), la base de ce système de traitement du langage et surtout le but premier de cet recherche puisque c'est elle que nous utilisons dans le processus d'apprentissage. Nous avons nommé cette fonction $Anti$, elle se définit comme:

$$\vartheta^2 \rightarrow \vartheta : \qquad X, C \rightarrow Z = Anti(X, C)$$

où ϑ représente l'ensemble des vecteurs conceptuels.

5.2. Items sans antonymes

Certains items n'ont pas d'antonymes. C'est le cas, par exemple, des objets matériels comme ‹*voiture*›, ‹*bouteille*›, ‹*bateau*›, etc. Nous devons donc nous poser la question de la continuité des fonctions d'antonymie dans le domaine des vecteurs conceptuels. Comment définir l'antonyme d'un item lexical qui n'a pas d'antonyme? Nous pouvions considérer que l'antonyme d'un item non opposable est $\vec{0}$ (le vecteur nul) ou considérer l'objet comme un point fixe de la fonction. En d'autre termes, que l'antonyme d'un item lexical qui ne possède pas d'antonyme est l'item lexical lui-même. La première approche ne semble pas pertinente. En terme de linguistique, cela reviendrait à considérer que le contraire d'un item non opposable est l'idée vide. Cela n'est pas acceptable avec la méthode de construction que nous définirons par la suite. En effet, si nous cherchions l'antonyme d'une ‹*ferrari*›, qui serait une *AUTOMOBILE*, *ROUGE* et *RAPIDE* nous ne voulons pas avoir une ‹*tortue*› (*VERTE* et *LENTE*) mais plutôt une *AUTOMOBILE*, *VERTE* et *LENTE*, une ‹*deux-chevaux*› par exemple. Elle offre l'avantage que les points fixes peuvent être considérés comme étant sur l'axe de symétrie, ce qui est compatible avec l'approche théorique précédente. Dans la suite de notre exposé, nous parlerons indifféremment d'items lexical qui ne possède pas d'antonymes ou d'items lexical qui est son propre antonyme.

Comme les fonctions de synonymies (Lafourcade and Prince, 2001), les diverses fonctions $Anti$ sont dépendantes du contexte mais, contrairement à elles, elles ne peuvent pas être indépendantes de l'organisation des concepts. Elles nécessitent d'identifier pour chaque concept et pour chaque contexte un vecteur qui sera considéré comme son opposé. Il faut donc construire une liste de triplets $\langle concept, contexte, vecteur \rangle$ appellé listes d'antonymes. Il est important de noter que cette liste est différente pour chaque type d'antonymie. Il suffit donc de dresser autant de listes que de types d'antonymie examinés.

5.3. Listes d'antonymes

5.3.1. Construction de la liste d'antonymes

Ces vecteurs antonymes sont construits, manuellement, uniquement à partir de vecteurs conceptuels de concepts générateurs (y compris le vecteur lui-même le cas échéant puisque comme nous l'avons déjà défini) l'antonyme d'un item lexical non opposable est lui même. Ainsi, nous pouvons avoir par exemple:

$AntiC(\textit{EXISTENCE}, V) = V(\textit{INEXISTENCE}) \quad \forall V$ $\qquad$ $AntiC(\textit{AGITATION}, V) = V(\textit{INERTIE}) \oplus V(\textit{REPOS}) \quad \forall V$
$AntiC(\textit{INEXISTENCE}, V) = V(\textit{EXISTENCE}) \quad \forall V$ $\qquad$ $AntiC(\textit{JOUET}, V) = V(\textit{JOUET}) \quad \forall V$

Comme nous l'avons vu dans le partie 2, des items lexicaux peuvent avoir, suivant le contexte, un antonyme différent. Bien qu'ils ne soient censés pas être polysémiques, il en est de même pour les concepts. Ainsi, *DESTRUCTION* peut avoir comme antonyme *PRÉSERVATION, CONSTRUCTION, RÉPARATION* ou *PROTECTION*. Nous avons ainsi défini pour chacun d'eux un vecteur conceptuel qui permettra la sélection de l'antonyme le mieux adapté à la situation.

concept	contexte	vecteurs constituant le vecteur antonyme.
EXISTENCE	$\forall V$	$V(\textit{INEXISTENCE})$
INEXISTENCE	$\forall V$	$V(\textit{EXISTENCE})$
AMOUR	$\forall V$	$V(\textit{DÉSACCORD}) \oplus V(\textit{AVERSION}) \oplus V(\textit{INIMITIÉ})$
DÉSORDRE	$V(\textit{ORDRE}) \oplus V(\textit{DÉSORDRE})$	$V(\textit{ORDRE})$
DÉSORDRE	$V(\textit{ORDRE}) \oplus V(\textit{CLASSIFICATION})$	$V(\textit{CLASSIFICATION})$
DÉSORDRE	$V(\textit{ORGANISATION}) \oplus V(\textit{DÉSORGANISATION})$	$V(\textit{ORGANISATION})$

Par exemple, le concept *EXISTENCE* a pour antonyme le vecteur *INEXISTENCE* quelque soit le contexte. Le concept *DÉSORDRE* a pour antonyme le vecteur *ORDRE* dans le contexte constitué des vecteurs *ORDRE* et *DÉSORDRE*.

5.3.2. Fonction $AntiC$

La fonction $AntiC$ renvoie en fonction d'un concept C_i et d'un contexte $V_{contexte}$ le vecteur considéré comme le vecteur antonyme dans une liste d'antonyme.

$$AntiC(C_i, V_{contexte})$$

Cette fonction se traduit donc par une simple exploration de la liste d'antonymes.

5.4. Construction du vecteur antonyme: fonction Anti

5.4.1. Définitions

Nous pouvons définir mathématiquement la fonction d'antonymie relative:

$$Anti_R : X, C \rightarrow Z = Anti_R(X, C)$$

La fonction d'antonymie absolue $Anti_A$ peut être définie mathématiquement comme:

$$Anti_A : X \rightarrow Z = Anti_A(X) = Anti_R(X, X)$$

Nous allons maintenant expliciter la fonction $AntiLex_S$ en montrant comment à partir de deux vecteurs conceptuels, un pour l'item lexical dont nous voulons l'antonyme, l'autre pour le contexte, nous construisons le vecteur conceptuel opposé.

5.4.2. Construction du vecteur conceptuel antonyme

L'idée est d'insister sur les notions saillantes dans V_{item} et V_c. Si ces notions peuvent être opposées, alors l'antonyme doit posséder les idées inverses dans la même proportion. La fonction d'antonymie est définie comme suit:

$$Anti_R(V_{item}, V_c) = \bigoplus_{i=1}^{N} P_i \times AntiC(C_i, V_c)$$

avec

$$P_i = V_{item_i}^{1+CV(V_{item})} \times max(V_{item_i}, V_{c_i})$$

Le poids P a été défini empiriquement à la suite d'expérimentations. Clairement, la fonction ne pouvait pas être symétrique. Nous ne devons pas avoir $Anti_R$(V(‹*chaud*›),V(‹*température*›)) = $Anti_R$(V(‹*température*›),V(‹*chaud*›)). La puissance $1 + CV(V_item)$ a donc été introduite pour insister d'avantage sur les idées présentent dans le vecteur que nous voulions opposer. Nous avons aussi remarqué que plus un vecteur était conceptuel (proche du vecteur d'un concept) plus il était intéressant d'augmenter cette puissance. C'est la raison pour laquelle cette puissance comprend le coefficient de variation[4] qui est un bon indice de la "conceptualité". Enfin, nous avons introduit la fonction *max* afin de considérer les idées de l'item même si celles-ci ne sont pas présentes dans le référent. Par exemple, si l'on cherche l'antonyme de ‹*froid*› dans le contexte de ‹*température*›, le poids de ‹*froid*› doit être important même s'il n'est pas présent dans le vecteur représentant ‹*température*›. Nous exposons dans 5.5.2 des exemples calculés à partir de ces formules. Maintenant, nous allons présenter les transitions entre les items lexicaux et les vecteurs.

5.5. Items lexicaux et vecteurs: problèmes et solutions

L'objectif des fonctions $AntiLex$ est, à partir d'un item lexical, de donner son ou ses items antonymes. Il faut utiliser la fonction $Anti$ définie en 5.4. Nous devons donc utiliser des outils permettant le passage des items lexicaux vers les vecteurs conceptuels. Ce passage pose de nombreux problèmes dû à la polysémie c'est à dire choisir la bonne correspondance entre un item lexical et un vecteur conceptuel, choisir le bon sens du mot.

5.5.1. Passage items lexicaux → Vecteurs conceptuels

Il s'agit du passage 1 de la figure (1). Nous l'avons vu, notre définition de l'antonymie est relative à un contexte. Dans certains cas, ce contexte seul ne suffit pas à déterminer un axe de symétrie pour l'antonymie. Rappelons l'exemple de l'item lexical ‹*père*› qui dans le contexte de ‹*famille*› est opposable à ‹*mère*› ou ‹*enfant*›.

4. Le coefficient de variation est donnée par la formule $\frac{EC(V)}{\mu(V)}$ avec EC(V) l'écart type du vecteur V et $\mu(V)$ la moyenne arithmétique des composantes de V.

La fonction $AntiLex_S$ prend trois items lexicaux en arguments: l'item lexical *item* dont on cherche l'antonyme et l'item lexical *contexte* qui définie le contexte qui nous permet de sélectionner le sens de *item* plus particulièrement recherché. Par exemple, le mot ‹*femme*› peut avoir comme acceptation épouse ou bien personne du sexe féminin. Dans le premier cas, un antonyme pourrait être ‹*mari*›, dans le second ‹*homme*›. Le troisième item sert d'axe de symétrie. Pour réaliser la sélection du sens de l'item et, s'il est différent du contexte, réaliser la sélection du sens du référent: nous utilisons une méthode de contextualisation forte qui à partir, d'un item, de ses définitions issues de dictionnaires et d'un vecteur conceptuel représentant le contexte renvoie un vecteur où certains sens sont privilégiés au détriment d'autres en fonction de ce contexte. De la même manière, le vecteur référent est aussi contextualisé.

Cette mise en contexte montre les limites de la fonction $Anti_R$ d'antonymie absolue. Dans ce cas, aucune contextualisation n'est réalisée donc la fonction d'antonymie se fera sur le vecteur brut du mot. Cela augmentera donc le bruit et donc aura moins de chance de fournir n résultat pertinent. Ce problème n'est pas spécifique à la recherche de l'antonymie, il est important d'essayer de fournir un contexte.

5.5.2. Passage Vecteurs conceptuels → items lexicaux

Il s'agit du passage 1 de la figure (1). Ce passage inverse du précédant est, a priori, plus aisé. Il suffit de comparer le vecteur conceptuel antonyme V_{ant} aux vecteurs conceptuels de la base. Les plus proches, au sens de la distance angulaire, c'est à dire ceux qui sont les plus synonymes de V_{ant}, seront en antonymie thématique de V_{item} (Lafourcade and Prince, 2001). Avec notre méthode, nous obtenons, par exemple,

- $AntiLex_R$(‹*vie*›,‹*existence*›) = $\mathcal{V}(Anti_R$(‹*vie*›, ‹*mort*› et ‹*vie*›)) = (‹*mort*›⌈*0.3367*⌋) (‹*être guéri de tous les maux*›⌈*0.3573*⌋) (*MORT*⌈*0.3573*⌋) (‹*haschischin*›⌈*0.3672*⌋) (‹*assassin*›⌈*0.3675*⌋) (‹*assassineur*›⌈*0.3774*⌋) (*C3:LES ÂGES DE LA VIE*⌈*0.481*⌋) (‹*tyrannicide*›⌈*0.5161*⌋) (‹*assassiner*›⌈*0.5797*⌋)
- $AntiLex_A$(*VIE*) = $\mathcal{V}(Anti_A$(*VIE*)) = (*MORT*⌈*0.0344*⌋) (‹*être guéri de tous les maux*›⌈*0.0344*⌋) (‹*mort*›⌈*0.4272*⌋) (*C3:LES AGES DE LA VIE*⌈*0.551*⌋) (‹*haschischin*›⌈*0.568*⌋) (‹*assassin*›⌈*0.5683*⌋) (‹*assassineur*›⌈*0.5885*⌋) (‹*tyrannicide*›⌈*0.6991*⌋) (*C2:L'ÊTRE HUMAIN*⌈ *0.7357*⌋) (‹*assassiner*›⌈*0.7485*⌋)

L'item lexical le plus proche du vecteur antonyme du concept *VIE* est *MORT*, ce qui est très acceptable. On peut s'apercevoir que la distance entre le vecteur antonyme calculé et le vecteur antonyme réel n'est pas nulle. Cela vient du fait que la méthode n'est pas et ne peut pas être exacte. L'objectif de nos fonctions est de construire le meilleur vecteur antonyme possible. Pour autant, si la distance entre le vecteur calculé et le vecteur de l'antonyme n'est pas nulle, il ne faut pas considérer que la méthode de calcul est mauvaise. Il faut considérer que le vecteur de l'antonyme doit être réappris grâce à la fonction d'antonymie. Plus le mot est polysémique, plus ce phénomène est fréquent comme en témoigne l'exemple $AntiLex_R$(‹*vie*›, ‹*existence*›) où l'antonyme est à 0.4 radians (23°) du vecteur calculé.

On ne peut considérer comme antonyme, même si la mesure de potentiel d'antonymie est correcte, l'item lexical le plus proche du vecteur antonyme. Un antonyme doit avoir la même catégorie morphologique que ‹*item*› (le résultat le plus pertinent pour un verbe sera de rendre un verbe). Nous avons présenté une méthode de construction des vecteurs conceptuels. Cette dernière permet de déterminer si un vecteur est susceptible de posséder un antonyme.

6. Mesure d'évaluation de l'antonymie

Il semble pertinent de savoir si deux items lexicaux peuvent être l'antonyme l'un de l'autre afin de posséder un outil comparable à la synonymie relative (Lafourcade and Prince, 2001).

Nous avons donc créé une mesure d'évaluation de l'antonymie. Soient les vecteurs A et B. La question est de savoir si on peut dire s'ils sont antonymes dans le contexte C. La distance d'antonymie $Manti_{Eval}$ est la mesure de l'angle formé par la somme par les vecteurs A et B et la somme de leur opposés $Anti_{c_R}(A, C)$ et $Anti_{c_R}(B, C)$. Soit:

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

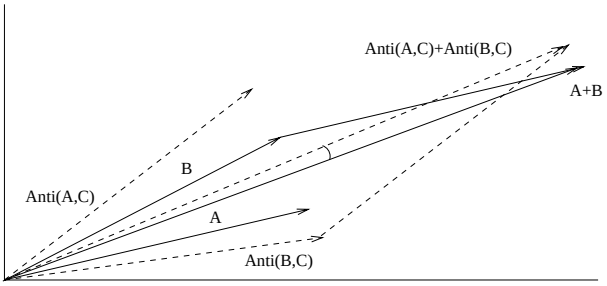

FIG. 2 – *Représentation géométrique en 2D de la mesure d'évaluation de l'antonymie*

La mesure d'antonymie n'est pas une distance. Ce n'est qu'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 $Mant = \pi/2$ l'un de l'autre n'ont aucune idée en commun[5]. Nous pouvons plutôt voir ici l'illustration que deux antonymes ont certaines idées en commun, celles qui ne sont pas 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 le fait que ces deux items lexicaux n'ont que peu d'idées en commun, 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 certaine idées!

6.1. Exemples

Dans les exemples qui suivent, le contexte est constitué par la somme des vecteurs des deux items.

$Manti_{Eval}$(*EXISTENCE,INEXISTENCE*)	$= 0.03$	$Manti_{Eval}$(‹*existence*›, ‹*automobile*›)	$= 1.06$
$Manti_{Eval}$(‹*existence*›,‹*inexistence*›)	$= 0.44$	$Manti_{Eval}$(*AUTOMOBILE, AUTOMOBILE*)	$= 0.006$
$Manti_{Eval}$(*EXISTENCE, AUTOMOBILE*)	$= 1.45$	$Manti_{Eval}$(‹*automobile*›, ‹*automobile*›)	$= 0.407$

Les exemples ci-dessus illustrent bien ce que nous disions auparavant. Les concepts *EXISTENCE* et *INEXISTENCE* sont très fortement antonymes en antonymie complémentaire. L'effet de la polysémie explique que les items ‹*existence*› et ‹*inexistence*› soient moins antonymes que les concepts. En antonymie complémentaire, *AUTOMOBILE* est son propre antonyme. La mesure de l'antonymie entre *AUTOMOBILE* et *EXISTENCE* est un exemple de notre remarque précédente sur les vecteurs qui

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

ne partagent que peu d'idées. Aux alentours de $\frac{\pi}{2}$, cette mesure se comporte comme la distance angulaire . D'ailleurs, nous avons $D_A(existence, automobile) = 1.464$. On pourait envisager d'utiliser cette fonction pour chercher, dans le lexique conceptuel, les meilleurs antonymes mais le coût en temps (environ une minute pour une mesure sur un P4 1.3Ghz) reste prohibitif.

7. Action sur l'apprentissage

Cette fonction d'antonymie a été intégrée au processus d'apprentissage associé à la création du lexique conceptuel. Cela a bien sûr une incidence sur les résultats de la mesure d'antonymie.

$Manti_{Eval}$(EXISTENCE,INEXISTENCE)	$= 0.03$	$Manti_{Eval}$(‹existence›, ‹automobile›)	$= 1.1$
$Manti_{Eval}$(‹existence›,‹inexistence›)	$= 0.33$	$Manti_{Eval}$(AUTOMOBILE, AUTOMOBILE)	$= 0.006$
$Manti_{Eval}$(EXISTENCE, AUTOMOBILE)	$= 1.45$	$Manti_{Eval}$(‹automobile›, ‹automobile›)	$= 0.30$

Si l'incidence est nulle sur les concepts puisque les vecteurs correspondant ne sont en aucun cas modifiés par l'apprentissage, en revanche, la mesure d'évaluation de l'antonymie sur les items est meilleure. Notre exemple, aprés intégration de la fonction d'intégration, montre que les vecteurs ‹*existence*› et ‹*inexistence*› ont étés modifiés largement. Maintenant, les deux items sont considérés comme "plus antonymes" que précédement. Les définitions se sont donc affinées et la base de vecteurs est maintenant d'autant plus cohérente. Bien entendu, on peut tester de tels résultats sur les 62000 entrées du lexique qui ont étés modifiées d'une façon plus ou directe par la fonction d'antonymie.

8. Conclusions et perspectives

Cet article présente une modélisation de l'antonymie à l'aide de vecteurs conceptuels. Dans le cadre du TALN, l'antonymie est un aspect fondammental dès que l'on cherche à évaluer et représenter la sémantique de segments textuels. Les applications majeures en sont l'analyse thématique de textes et la construction de grandes bases lexicales. Nous nous sommes basés sur une théorie linguistique calculable ainsi que vectoriellement représentable et nous avons mené notre analyse dans la perspective de la symétrie. Ce travail préliminaire nous a permis d'exprimer en termes de vecteurs conceptuels diverses fonctions d'antonymie, à savoir les antonymies complémentaire, scalaire et duale. Ces fonctions permettent, à partir d'un vecteur conceptuel et d'informations contextuelles, de calculer un vecteur antonyme. Des extensions ont aussi été réalisées pour que ces fonctions soient définies et utilisables à partir des items lexicaux. Des mesures ont été introduites: l'une permet d'évaluer si un vecteur est succeptible de posséder un antonyme, l'autre d'apprécier si deux items peuvent êtres antonymes l'un de l'autre. Ces mesures d'antonymie sont à la base de la détection des phénomènes d'opposition dans les textes. La construction de vecteurs, elle, est nécessaire pour la sélection des items lexicaux contraires dans le cas de la génération de textes.

De nombreuses améliorations sont possibles à commencer par la révision des listes d'antonymes. Elles n'ont été actuellement construites que par un groupe réduit de personnes et il est souhaitable que d'autres, en particulier des linguistes, les affinent. L'utilisation de ces fonctions mettra certainement en évidence des problèmes que nous n'avons pas encore rencontrés. Des corrections devront probablement être apportées à nos fonctions et mesures pour quelles soient plus précises. Nous pensons en particulier à la mesure de potentiel d'antonymie dont les valeurs autour de 0 ne sont pas exploitables. Une piste possible est celle de l'automatisation de ces ajustements par apprentissage sur des corpus.

Références

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

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

Lafourcade M. and Prince V. (2001). Synonymies et vecteurs conceptuels. In *actes de TALN'2001*, Tours, France.

Larousse (1991). *Grand Larousse Universel.* Larousse.

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

Larousse (2001). *Le Petit Larousse Illustré 2001*. Larousse.

Mel'čuk I., Clas A., and Polguère A. (1995). *Introduction à la lexicologie explicative et combinatoire.* Duculot.

Morin E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus techniques.* Université de Nantes.

Salton G. and McGill M. (1983). *Introduction to Modern Information Retrieval*. McGrawHill.

Schwab D. (2001). *Vecteurs conceptuels et fonctions lexicales : application à l'antonymie*. université Montpellier II.

Véronis J. and Ide N. (1992). A feature-based model for lexical databases. In *actes de COLING 92.*