TALN 2004, Fès, 19–21 avril 2004

Désambiguïsation par proximité structurelle

Bruno Gaume (1), Nabil Hathout (2) & Philippe Muller (1)
(1) IRIT – CNRS, UPS & INPT
{gaume,muller}@irit.fr
(2) ERSS – CNRS & UTM
hathout@univ-tlse2.fr

Résumé - Abstract

L'article présente une méthode de désambiguïsation dans laquelle le sens est déterminé en utilisant un dictionnaire. La méthode est basée sur un algorithme qui calcule une distance « sémantique » entre les mots du dictionnaire en prenant en compte la topologie complète du dictionnaire, vu comme un graphe sur ses entrées. Nous l'avons testée sur la désambiguïsation des définitions du dictionnaire elles-mêmes. L'article présente des résultats préliminaires, qui sont très encourageants pour une méthode ne nécessitant pas de corpus annoté.

This paper presents a disambiguation method in which word senses are determined using a dictionary. We use a semantic proximity measure between words in the dictionary, taking into account the whole topology of the dictionary, seen as a graph on its entries. We have tested the method on the problem of disambiguation of the dictionary entries themselves, with promising results considering we do not use any prior annotated data.

Mots-clefs – Keywords

Désambiguïsation sémantique, réseaux petits mondes hiérarchiques, dictionnaires.
Word sense desambiguation, hierarchical small words, dictionaries.

1 Introduction

De nombreuses tâches impliquant le traitement de données en langue naturelle sont rendues difficiles par l'existence de sens différents pour un même item lexical : traduction automatique, recherche de documents ou extraction d'informations. Ce problème, très ancien en TAL, est loin d'être résolu, et l'évaluation de ses méthodes est difficile et relativement récente, pour des raisons présentées notamment dans (Resnik & Yarowsky, 2000). On peut distinguer plusieurs familles d'approches[1], selon que le sens d'un mot en contexte est déterminé en apprenant automatiquement les caractéristiques du contexte qui détermine ce sens (de façon supervisée, ou non supervisée, quand l'étude des contextes sert elle-même à dégager des familles d'usage) ou bien que le sens soit déterminé en utilisant des ressources lexicales « extérieures » : dictionnaires, thésaurus. Le premier type d'approche nécessite des données volumineuses difficiles à annoter (pour les approches supervisées ; les approches non supervisées sont par ailleurs sensibles au corpus choisi, qui doit être représentatif). Le deuxième type d'approche tente d'utiliser la connaissance lexicale rassemblée dans les dictionnaires, les thésaurus (WordNet, par exemple), avec une longue tradition, (Lesk, 1986; Banerjee & Pedersen, 2003) et des résultats mitigés. Dans tous les cas, on cherche à établir une relation de distance entre mots, susceptible de déterminer un sens en contexte. Dans le cas des dictionnaires, les seules méthodes ayant présenté des résultats chiffrés se concentrent seulement sur les mots qui apparaissent dans la définition d'un mot cible[2].

Nous présentons ici un algorithme qui utilise un dictionnaire comme source d'information sur les relations entre items lexicaux (*cf.* section 3). L'algorithme calcule une distance « sémantique » entre les mots du dictionnaire en prenant en compte la topologie complète du dictionnaire, ce qui lui donne une plus grande robustesse. Nous avons commencé à tester cette approche sur la désambiguïsation des définitions du dictionnaire elles-mêmes (section 2), mais nous montrons pourquoi cette méthode est plus générale. La section 6 présente nos résultats préliminaires, qui sont très encourageants pour une méthode ne nécessitant pas de corpus annoté (en dehors de l'évaluation), et qui comporte de nombreux paramètres d'ajustement.

2 Le graphe du dictionnaire

L'idée de base de notre méthode est de considérer qu'un dictionnaire est un graphe non orienté dont les mots sont les sommets et tel qu'il existe un arc entre deux sommets si l'un apparaît dans la définition de l'autre. Plus précisément, le graphe du dictionnaire encode deux types d'informations lexicographiques : les définitions qui décrivent les différentes acceptions de chaque vedette au moyen de séquences langagières ; la structure des articles qui organise ces sous-sens[3]. Deux types de sommets sont ainsi nécessaires : les sommets-w qui représentent les mots qui apparaissent dans les définissants, et les sommets-Δ qui correspondent aux sous-sens des vedettes. La construction du graphe se fait en trois temps :

1. Pour chaque vedette, on crée un sommet-Δ qui correspond à l'article entier et autant de sommets-Δ qu'il y a de sous-sens pour lesquels il existe un définissant. On crée un

[1] On peut se référer au numéro spécial de *Computationnal Linguistics* de 1998 et son introduction (Ide & Véronis, 1998) ; cf. aussi (Manning & Schütze, 1999, chap. 7).

[2] On peut citer aussi les propositions non quantifi ées de (H.Kozima & Furugori, 1993).

[3] Nous adoptons ici la terminologie de (Martin, 1983) et (Henry, 1996)

arc entre chaque sommet-Δ et les sommets-Δ qui représentent des sous-sens de niveau immédiatement inférieur.

2. Pour chaque mot qui apparaît dans un définissant du dictionnaire, on crée un sommet-w. On crée un arc entre chaque couple de sommets $\langle w, \Delta \rangle$ si le mot représenté par le sommet-w apparaît dans le définissant du sous-sens correspondant au sommet-Δ.
3. On crée un arc entre chaque couple de sommets $\langle w, \Delta \rangle$ si le sommet-Δ représente l'article dont la vedette est le mot correspondant au sommet-w.

Considérons, à titre d'exemple, l'article de « daim, n. m. » (issue du dictionnaire *Le Robert*) :

1. Mammifère ruminant ongulé.
2. [a] Peau préparée de cet animal.
   [b] Cuir suédé (veau retourné).
3. Corne de daim [...]
4. Bellâtre.

Le graphe contiendra un premier sommet (appelons le Δ_0) qui représente l'article dans sa totalité. Δ_0 est relié par un arc à chacun des sommets Δ_1, Δ_2, $\Delta_3$et Δ_4 qui représentent respectivement les sous-sens 1., 2., 3. et 4. À son tour Δ_2 est connecté à deux sommets $\Delta_{2.1}$ et $\Delta_{2.2}$ correspondant aux sous-sens 2.[a] et 2.[b]. Le graphe contient ensuite des arcs qui vont de Δ_1 vers trois sommets w_1, w_2 et w_3 qui représentent les mots *mammifère*, *ruminant* et *ongulé*. Enfin, il y aura un arc entre w_1 et le sommet-Δ qui représente l'article de « mammifère, adj. et n. », etc.

L'expérience que nous présentons a été réalisée au moyen d'un graphe construit à partir de définitions issues du dictionnaire *Le Robert*. Ce graphe est restreint aux seuls substantifs : il n'inclut que des définissants de vedettes nominales dans lesquelles n'ont été conservées que les occurrences nominales.

Dans les articles, les sous-sens s'inscrivent dans des structures hiérarchiques qui peuvent comporter jusqu'à cinq niveaux : ^{1}, ^{2}, ^{3}... pour les homographes ; I, II, III... ; A, B, C... ; 1, 2, 3... et a, b, c... pour les acceptions. Les positions de ces sous-sens peuvent ainsi être représentées de manière uniforme au moyen de séquences de cinq nombres correspondant aux cinq niveaux. Par exemple, le sous-sens 2.[a] de l'article de *daim* est décrit par 0_0_0_2_1 (les trois premiers niveaux n'étant pas utilisés, ils sont représentés par des zéros).

3 PROX : une méthode pour la mesure de similarité lexicale

PROX est une méthode stochastique pour l'étude de la structure des réseaux petits mondes hiérarchiques (voir section suivante). Cette méthode consiste à transformer un graphe en une chaîne de Markov dont les états sont les sommets du graphe en question et ses arêtes les transitions possibles : une particule en partant à l'instant $t = 0$ d'un sommet s_0, se déplace en un pas sur s_1 l'un des voisins de s_0 sélectionné aléatoirement ; la particule se déplace alors à nouveau en un pas sur s_2, l'un des voisins de s_1 sélectionné aléatoirement etc. Si au t-ième pas la particule est sur le sommet s_t elle se déplace alors en un pas sur le sommet s_{t+1} qui est sélectionné aléatoirement parmi les voisins de s_t tous équiprobables. Une trajectoire s_1, s_2, ...

s_t, ... ainsi sélectionnée est une « balade » aléatoire sur le graphe, et ce sont les dynamiques de ces trajectoires qui nous donnent les propriétés structurelles des graphes étudiés.

Posons PROX(G, i, r, s) la probabilité qu'en partant à l'instant $t = 0$ du sommet r la particule soit à l'instant $t = i$ sur le sommet s :

1. Un graphe non orienté $G = (V, E)$ est la donnée d'un ensemble non vide fini V de sommets, et d'un ensemble E de paires de sommets formant des arêtes. Si l'arête $\{r, s\} \in E$ on dit que les sommets r et s sont voisins, le nombre de voisins d'un sommet r est $d(r)$ son degré d'incidence ;

2. Soit un Graphe à n sommets $G = (V, E)$, on notera $[G]$ la Matrice carrée $n \times n$ telle que pour tout $r, s \in V \times V$, $[G]_{r,s} = 1$ si $\{r, s\} \in E$ et $[G]_{r,s} = 0$ si $\{r, s\} \notin E$; On appellera $[G]$ la matrice d'adjacence de G. C'est-à-dire que $[G]_{r,s}$ (la valeur située à la r-ième ligne et la s-ième colone de la matrice $[G]$) est égale à 1 s'il existe une arête entre les sommet r et s, 0 sinon.

3. Soit $G = (V, E)$ un graphe à n sommets. Posons $[\hat{G}]$ la matrice $n \times n$ de transition de la chaîne de Markov homogène dont les états sont les sommets du graphe en question telle que la probabilité de passer d'un sommet $r \in V$ à l'instant i vers un sommet $s \in V$ à l'instant $i + 1$ est égale à :

   $[\hat{G}]_{r,s}$=0 si $\{r, s\} \notin E$ (s n'est pas un voisin de r) $[\hat{G}] = 1/d(r)$ si $\{r, s\} \in E$ (s est un des $d(r)$ voisins de r qui sont tous équiprobables)

   Nous dirons que $[\hat{G}]$ est la matrice Markovienne du graphe G et que G est le graphe des transitions possibles de cette chaîne de Markov.

4. Soit $G = (V, E)$ un graphe réflexif à n sommets et $[\hat{G}]$ sa matrice Markovienne, pour tout $r, s \in V \times V$, on a donc :

   PROX$(G, i, r, s) = [\hat{G}^i]_{r,s}$

   où A^i est la matrice A multipliée i fois par elle-même.

C'est-à-dire que pour tout r, s, PROX(G, i, r, s) est la probabilité que la particule en partant du sommet r à l'instant $t = 0$ soit à l'instant $t = i$ sur le sommet s quand elle se déplace aléatoirement de sommet en sommet dans le graphe en empruntant les arêtes du graphe.

Si PROX$(G, i, r, s) >$ PROX(G, i, r, u) cela veut donc dire que dans sa trajectoire la particule en partant du sommet r, à plus de chance d'être à l'instant i sur le sommet s que sur le sommet u, et c'est la structure du graphe qui détermine ces probabilités.

PROX construit ainsi une mesure de similarité entre sommets d'un graphe, en « rapprochant » les sommets d'une même zone dense[4] en arêtes, ce qui permet d'envisager une exploitation originale et novatrice des dictionnaires électroniques (Gaume *et al.*, 2002) avec un outil de visualisation du sens (Gaume & Ferré, 2004), mais aussi de construire des outils pour le TAL, par exemple la désambiguïsation des entrées dans un dictionnaire. Pour une présentation détaillée de PROX voir (Gaume, à paraître).

[4]En effet plus il existe un grand nombre de chemins courts entre deux sommets r et s, plus la probabilité PROX(G, i, r, s) que la particule en partant du sommet r à l'instant $t = 0$ soit à l'instant $t = i$ sur le sommet s est grande.

4 Les graphes de dictionnaires sont des petits mondes hiérarchiques

Des recherches récentes en théorie des graphes ont mis au jour un ensemble de caractéristiques statistiques que partagent la plupart des grands graphes de terrain ; ces caractéristiques définissent la classe des graphes de type « réseaux petits mondes hiérarchiques » (RPMH ; en anglais *hierarchical small world*) (Watts & Strogatz, 1998; Newman, 2003). Les RPMH présentent quatre propriétés fondamentales :

D : ils sont peu denses, c'est-à-dire qu'ils ont relativement peu d'arêtes au regard du nombre de leurs sommets ;

L : la moyenne des plus courts chemins entre les sommets est petite ;

C : le taux de clustering ou d'agrégation, est défini de la manière suivante : Supposons qu'un sommet S ait Ks voisins, alors il y a Ks(Ks-1)/2 arêtes au maximum qui peuvent exister entre ses Ks voisins (ce qui arrive quand chacun des voisins de S est connecté à tous les autres voisins de S). Soit As le nombre d'arêtes qu'il y a entre les voisins de S (ce nombre est donc nécessairement plus petit ou égal à Ks(Ks-1)/2). Posons Cs= As/(Ks(Ks-1)/2) qui est donc pour tout sommet S inférieur ou égal à un. Le C d'un graphe est la moyenne des Cs sur ses sommets. Le C d'un graphe est donc toujours compris entre 0 et 1. Plus le C d'un graphe est proche de 1, plus il forme des agrégats ou clusters (des zones denses en arêtes). Dans un RPMH le C est fort, deux voisins d'un même sommet ont tendance à être connectés entre eux par une arête (« mes amis sont amis entre eux »). Par exemple, sur Internet[5], deux pages qui sont liées à une même page ont une probabilité relativement élevée d'inclure des liens l'une vers l'autre ;

I : la distribution des degrés d'incidence des sommets suit une loi de puissance (*power law*) : certains nœuds très peu nombreux ont beaucoup plus de voisins que d'autres plus nombreux, eux-mêmes ayant plus de voisins que d'autres qui eux-mêmes... La probabilité $P(k)$ qu'un sommet du graphe considéré ait k voisins décroît comme une loi de puissance $P(k) = k^{-\lambda}$.

Le tableau 1 présente une comparaison des RPMH avec d'autres types de graphes pour ces différentes caractéristiques : des graphes aléatoires (construit en partant d'un ensemble de sommets isolés, puis en ajoutant aléatoirement un certains nombre déterminé d'arêtes entre ses sommets), et des graphes réguliers (des graphes classiquement étudiés en théorie des graphes, dont tous les sommets ont le même degré d'incidence).

Les graphes d'origine linguistique et notamment ceux qui sont construits à partir de dictionnaires sont de type RPMH. Par exemple le graphe G1 des noms construit à partir du dictionnaire *Le Robert* (les sommets sont les entrées qui sont des noms, et il existe une arrête entre deux sommets si l'un est dans la définition de l'autre - on ne tient pas compte ici de la structure hiérarchique des définitions) est un RPMH typique (voir table 2) . Dans le graphe G2 qui est construit comme indiqué à la section 2, chaque sommet est remplacé par l'arbre reflétant la structure hiérarchique de l'entrée qui lui correspond, ce qui a pour conséquence d'affaiblir le C et d'allonger le L. Dans le tableau ci-dessous, * indique que les mesures sont calculées sur la plus grande partie connexe.

[5]Les sommets en sont les 800 millions de pages disponibles sur internet, et une arête est tracée entre A et B si un lien hypertexte vers la page B apparaît dans la page A ou si un lien hypertexte vers la page A apparaît dans la page B.

à densité égale	L : Moyenne des plus courts chemins	C : Taux de clustering	I : distribution des degrés d'incidences
Graphes aléatoires	L petit (chemins courts)	C petit (pas d'agrégats)	loi de Poisson
Graphes de terrain (RPMH)	**L petit (chemins courts)**	**C grand (des agrégats)**	**loi de puissance**
Graphes réguliers	L grand (chemins longs)	C grand (des agrégats)	constante

TAB. 1 – Comparaison de trois types de graphes en fonction des paramètres L, C et I.

Graphe	Nb. sommets	Nb. arcs	Nb. sommets*	Nb. Arcs*	Diamètre*	C*	L*
G1	51 559	392 161	51 511	392 142	7	0,182 9	3,32
G2	140 080	399 969	140 026	399 941	11	0,008 1	5,21

TAB. 2 – Quelques caractéristiques des graphes G1 et G2

Nous pensons que la nature hiérarchique des dictionnaires (distribution des degrés d'incidence des sommets en loi de puissance) est une conséquence du rôle de l'hyperonymie associée à la polysémie de certains sommets, alors que le fort C (existence de zones denses en arêtes) reflète le rôle de la cohyponymie[6] (Duvignau, 2002; Duvignau, 2003; Gaume *et al.*, 2002). Par exemple, le mot *corps* se trouve dans de nombreux définissants (*tête*, *chimie*, *peau*, *division*). De ce fait, le sommet *corps* a une forte incidence. D'autre part on constate qu'il existe de nombreux triangles par exemple : {*écorce*, *enveloppe*}, {*écorce*, *peau*}, {*peau*, *enveloppe*}, ce qui favorise les zones denses en arêtes et plus précisément un fort taux de clustering C. Ce sont ces zones denses en arêtes qui orientant la dynamique des trajectoires de la particule vont permettre la désambiguïsation.

5 Un algorithme de désambiguïsation basé sur PROX

Nous allons maintenant présenter une méthode pour désambiguïser une entrée de dictionnaire en utilisant la notion de distance sémantique introduite plus haut. On peut définir la tâche comme suit : on considère un lemme α qui apparaît dans la définition de l'un des sens d'un mot, considéré comme un nœud du graphe, β. Nous voulons donc associer α avec le sense le plus probable qu'il a dans ce contexte. Chaque entrée du dictionnaire est codé par un arbre de sous-sens dans le graphe du dictionnaire, avec une liste de nombres correspondants à chaque niveau de sous-sens caractéristique.

[6]Par exemple « enveloppe », « peau », « bogue », « écaille », « épluchure », « écorce », « vernis », « croûte », « enduit », « faux-semblant », « aspect », « apparence », « manteau », « fourrure », « toison », « pelure » sont rattachés à un même concept ENVELOPPE-APPARENCE constituant pour chacun d'entre eux, un noyau de sens commun. De tels mots constituent, de ce fait, des co-hyponymes, dont on peut distinguer deux types : - les co-hyponymes intra domaine : [« broue », « bogue », « cosse »,] ou [« pelage », « toison », « fourure »,] ou encore [« robe », « habit », « vêtement »,] qui relèvent d'un même domaine, à savoir respectivement dans ces exemples : VEGETAL ou ANIMAL ou HUMAIN. - les co-hyponymes inter-domaines : « écorce » et « pelage » sont des co-hyponymes inter-domaines car ils relèvent de domaines différents, respectivement le VEGETAL et l'ANIMAL. Le point commun de tous les hyponymes c'est leur potentialité à pouvoir exprimer la même idée en « intension ». C'est pourquoi ils peuvent être considérée comme co-hyponymes.

Soit $G = (V, E)$ un graphe à n sommets construit comme présenté section 2. L'algorithme suivant a été appliqué.

1. on supprime les voisins de β dans G $\forall x \in V [G]_{\beta,x} = [G]_{x,\beta} = 0$;
2. on calcule $[\hat{G}]^i$; nous avons pris $i = 6$ (cf. l'explication plus bas) :
3. soit L, le vecteur ligne de β alors $\forall k, L[k] = [\hat{G}]^i_{\beta,k}$;
4. Soit $F = \{x_1, x_2, ..., x_n\}$ les nœuds correspondant à tous les sous-sens de la définition de α.
   On prend alors $x_k = argmax_{x \in F}(L[x])$

Nous avons alors que x_k est le sous-sens le plus « proche » du nœud β, par rapport à la mesure Prox. Deux étapes demandent un peu plus d'explication :

1. les voisins sont supprimés pour ne pas laisser un biais favorable aux sous-sens de β, qui formeraient alors une sorte de cluster artificiel par rapport à la tâche donnée. Ainsi la « marche aléatoire » dans le graphe peut vraiment avoir lieu dans le graphe plus général des autres sens.
2. choisir une bonne valeur pour la longueur de la marche aléatoire n'est pas simple, et est le facteur essentiel de la réussite de la procédure. Si elle est trop petite, seules les relations locales vont apparaître (synonymes proches, etc) et ils peuvent ne pas apparaître dans les contextes à désambiguïser (c'est notamment le problème de la méthode de (Lesk, 1986)) ; si la valeur de i est trop grande par contre, les « distances » entre tous les mots tendent à converger vers une constante, faisant disparaître les différences. Cette valeur doit donc être reliée d'une façon ou d'une autre à la distance moyenne entre deux sens quelconques du graphe. Une hypothèse raisonable est donc de rester proche de cette valeur, et nous avons donc pris le nombre 6, la moyenne calculée étant de 5,21 (sur le graphe contenant tous les sous-sens, pas sur celui-ne contenant que les entrées, pour lequel $L = 3{,}3$) [7].

6 Évaluation

Pour chaque couple de sommets $(\alpha, \beta) \in V \times V$ tel que β représente un définissant Δ et α le lemme d'une forme qui apparaît dans Δ, l'algorithme précédent propose un sommet γ tel que γ appartient à la structure hiérarchique de l'article dont le mot-vedette est α et tel que γ permet d'identifier le sous-sens principal[8] de α qui sémantiquement est le plus proche de son occurrence dans Δ.

L'évaluation de la désambiguïsation sémantique a été réalisée comme suit : Nous avons sélectionné aléatoirement 27 définissants de substantifs dans le dictionnaire *Le Robert*. Deux personnes ont annoté sémantiquement les formes nominales qui y apparaissent. 82 triplets ont ainsi été constitués, dont il est resté 72 après avoir éliminé les mots ayant un seul sens dans le dictionnaire. Nous avons constaté que les désaccords entre les deux annotateurs ont été très rares et qu'un consensus a pu être trouvé rapidement dans les cas litigieux. Parallèlement, nous avons

[7] La valeur de L est calculée en appliquant une variante de l'algorithme de Dijkstra partant d'un nœud vers tous les autres, répétée pour chaque nœud du graphe.

[8] Le sous-sens principal correspond à la première sous-division hiérarchique d'une entrée, choisie parmi I, II ou III, ou bien A, B, ... suivant les cas, celà n'étant pas homogène dans le dictionnaire.

appliqué l'algorithme précédent aux 72 couples formés par les deux premiers éléments des triplets annotés. Nous avons ensuite comparé les résultats de l'algorithme avec les annotations manuelles. Ont été comptés comme corrects les solutions telles que le numéro d'homographie et le numéro de la première division hiérarchique sont identiques à ceux qui ont été proposés par les annotateurs. C'est le cas par exemple pour les couples des deux premières lignes du tableau suivant :

	β	α	γ	annotateurs
correct	bal#n._m.*0_0_0_3_0	lieu	**1_1**_0_3_0	**1_1**_0_1_0
correct	van#n._m.*2_0_0_0_0	voiture	**0_2**_0_0_0	**0_2**_0_3_0
erreur	phonétisme#n._m.*0_0_0_0_0	moyen	**1**_1_0_1_0	**2**_0_0_1_0
erreur	créativité#n._f.*0_0_0_0_0	pouvoir	**2_0_0_3**_0	**2_0_0_1**_0
erreur	acmé#n._m._ou_f.*0_0_0_1_0	phase	**0_0_0_1**_0	**0_0_0_4**_0

Pour les trois derniers couples, l'algorithme a proposé des solutions erronées : mauvais numéro d'homographe et / ou mauvais sous-sens principal.

Pour avoir une idée de la difficulté de la tâche, nous avons aussi calculé la moyenne des sous-sens principaux sur les entrées considérées, la moyenne du nombre d'homographes ayant la même catégorie grammaticale (nom commun) et la moyenne des sous-sens de niveau le plus fin des entrées considérées. Les résultats sont résumés dans la table 3. Le score de désambiguïsation

	hasard	algorithme	
homographes	0,5	0,8	(8/10)
polysémie principale	0,37	0,542	(39/72)
polysémie fine	0,125	0,292	(21/72)

TAB. 3 – Premiers résultats de l'évaluation de l'algorithme, avec une baseline aléatoire

des homographes n'est pas très significatif vu le petit nombre relevé dans les entrées choisies. Nous pouvons remarquer que les autres scores, sans être très bons, sont plutôt encourageants. Pour donner une idée de leur valeur, (Banerjee & Pedersen, 2003) applique des notions de distance lexicale variées issues de dictionnaire, appliquées à la désambiguisation (en anglais) de mots sélectionnés, avec des résultats qui vont de 0,2 à 0,4 par rapport aux sous-sens fournis par WordNet (et une moyenne de sous-sens par noms qui équivaudrait à 0,2 pour le score au hasard).

7 Conclusion

Nous avons présenté ici un algorithme donnant une mesure de similarité lexicale à partir d'un dictionnaire général. Cet algorithme est non supervisé. Il ne nécessite pas de corpus annoté et n'utilisant pas d'autres données qu'un dictionnaire général dont la couverture lexicale est la seule restriction sur le vocabulaire. La méthode donne des résultats prometteurs pour la désambiguïsation sur les noms seuls. Nous envisageons bien sûr d'étendre les tests à d'autres catégories grammaticales, mais aussi d'affiner la méthode pour les substantifs en considérant par exemple également les occurrences verbales dans les définissants des noms. Pour étendre cette méthode au cas général de la désambiguïsation, nous pensons par ailleurs considérer un

contexte qui contient un mot à désambiguïser comme une définition virtuelle que l'on ajouterait au graphe des mots pour appliquer ensuite exactement la même méthode. Nous envisageons également de réaliser des mesures plus fines des performances en tenant compte des degrés de confiance attribués à chaque candidat à la désambiguïsation (Resnik & Yarowsky, 2000).

Références

BANERJEE S. & PEDERSEN T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico.

DUVIGNAU K. (2002). *La métaphore berceau et enfant de la langue*. Thèse de doctorat, Université Toulouse - Le Mirail.

DUVIGNAU K. (2003). Métaphore verbale et approximation. *Revue d'Intelligence Artificielle*, 17(5/6), 869–881. Regards croisés sur l'analogie.

GAUME B. (à paraître). Balades aléatoires dans les petits mondes lexicaux. *I3 Information Interaction Intelligence*.

GAUME B., DUVIGNAU K., GASQUET O. & GINESTE M.-D. (2002). Forms of meaning, meaning of forms. *Journal of Experimental and Theoretical Artificial Intelligence*, 14(1), 61–74.

GAUME B. & FERRÉ L. (2004). Représentation de graphes par acp granulaire. In *Actes d'EGC 2004 : 4èmes journées d'Extraction et de Gestion des Connaissances*, Clermont-Ferrand.

HENRY F. (1996). Pour l'informatisation du TLF. In D. PIOTROWSKI, Ed., *Lexicographie et informatique. Autour de l'informatisation du* Trésor de la Langue Française, Paris: Didier Érudition.

H.KOZIMA & FURUGORI T. (1993). Similarity between words computed by spreading activation on an english dictionary. In *Proceedings of the conference of the European chapter of the ACL*, p. 232–239.

IDE N. & VÉRONIS J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1).

LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, p. 24–26, Toronto, Canada.

MANNING C. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

MARTIN R. (1983). *Pour une logique du sens*. Paris: Presses Universitaires de France.

NEWMAN M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, volume 45, 167–256.

RESNIK P. & YAROWSKY D. (2000). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2), 113–133.

WATTS D. & STROGATZ S. (1998). Collective dynamics of 'small-world' networks. *Nature*, (393), 440–442.