Quantité d'information échangée : une nouvelle mesure de la similarité des mots

Cédrick Fairon[1], Ngoc-Diep Ho[2]

[1]UCL – FLTR – CENTAL – 1348 Louvain-la-Neuve – Belgique
[2]UCL – FSA – INMA – 1348 Louvain-la-Neuve – Belgique
fairon@tedm.ucl.ac.be, ho@inma.ucl.ac.be

Abstract

There are a lot of methods for measuring semantic similarity between words that are based on different approaches. This document proposes a method based on the analyses of a dictionary. The definitions of words in the dictionary create a network. Its nodes are the headwords found in the dictionary and its edges represent relations between a headword and the words found in its definitions. The meaning of a word is defined by the *total quantity of information*, which each element of its definition contributes in. The similarity between two words is defined by the maximal *quantity of information exchanged* between them through the network.
In order to assess the performance, our measure of similarity will be compared with others measures and some applications based on our similarity will be also constructed.

Résumé

Il existe beaucoup de méthodes pour mesurer la similarité entre mots et ces méthodes se basent souvent sur des approches différentes. La recherche que nous présentons a pour but de proposer une nouvelle méthode basée sur l'analyse d'un dictionnaire. Les définitions du dictionnaire créent un réseau dont les nœuds sont les entrées lexicales du dictionnaire, les arcs sont des liens représentants la relation entre une entrée et les mots de ses définitions. Le sens d'un mot dépend de la *quantité totale* d'*information* que chaque mot dans sa définition va lui communiquer. La similarité entre 2 mots est définie par la *Quantité d'Information Echangée* (QIE) entre 2 mots, à travers le réseau.
Notre mesure de similarité sera comparée avec d'autres mesures et quelques applications basées sur cette mesure seront réalisées.

Mots-clés : similarité de mots, extraction de synonymes, filtre sémantique, flot maximal.

1. Introduction

Dans les dictionnaires explicatifs comme le *Petit Robert*, on trouve très souvent des synonymes ou des antonymes pour un mot quelconque. Par exemple, le mot « *maison* » et le mot « *logement* » sont synonymes, le mot « *maison* » et le mot « *abri* » sont aussi synonymes. Mais, comment peut-on dire que la connexion entre « *maison* » et « *logement* » est plus forte que celle entre « *maison* » et « *abri* » ? La réponse à cette question implique la notion de *similarité des mots* qui peut se représenter par une valeur scalaire qui définit comment 2 mots se relient. Plus concrètement, si la similarité entre le mot m_1 et le mot m_2 est quantifiée par $sim(m_1, m_2)$, on peut dire que « *maison* » est plus proche de « *logement* » que de « *abri* » si on a $sim(maison, logement) > sim(maison, abri)$ et vice versa.

La formalisation et la quantification de la similarité des mots ont été introduites depuis très longtemps. Cela remonte au moins à l'époque d'Aristote (384 – 322 B.C) (Budanisky, 1999), mais ces préoccupations n'avaient pas, jusqu'à il y a peu, trouvé beaucoup d'applications concrètes.

Dans le domaine du Traitement Automatique du Langage Naturel (TALN), les relations sémantiques comme la synonymie, l'antonymie, l'hyponymie, la méronymie, etc. sont des notions particulièrement importantes. Avec le développement de l'informatique et du Web, il y a chaque jour plus d'information disponible sur les pages du Web et sur les archives électroniques. C'est un avantage remarquable que l'on n'aurait jamais imaginé au début du siècle précédent. Cependant, pour les informaticiens, ça pose un problème pratique : comment trouver les informations utiles et gérer ces informations ? C'est la question qui a mené au développement du domaine de la *Recherche d'Information (RI)*. Jusqu'à présent, la plupart des méthodes utilisées dans les moteurs de recherche sont basées sur l'analyse statistique des occurrences des mots dans les documents. Cela marche bien dans beaucoup de cas, mais il y a encore des cas où ces méthodes ne sont pas satisfaisantes. Par exemple, quand on lance une recherche avec le mot clé : « *oiseau* », on ne veut pas seulement obtenir les documents contenant le mot « *oiseau* ». On attend aussi les documents qui contiennent les synonymes ou les mots étroitement liés avec le mot « *oiseau* » au niveau sémantique. Dans ce contexte, les *mesures de similarité* entre les mots sont particulièrement utiles.

Penchons-nous un instant sur la structure d'un dictionnaire explicatif. On peut constater en observant une entrée lexicale donnée que les mots trouvés dans sa définition jouent le rôle de *fournisseur* d'information. On remarque également que la plupart de ces mots seront définis par ailleurs dans d'autres notices où ils occuperont la position d'entrée lexicale et où ils seront cette fois en position de *récepteur* d'information. Dans ce contexte, la quantité d'information qu'un mot peut recevoir et fournir dépend donc de chaque mot et peut être calculée avec l'aide de la théorie de l'information. C'est sur cette constatation que nous nous fondons pour définir notre mesure de similarité basée sur la *Quantité d'Information Echangée* (QIE) et l'appliquer à *l'extraction automatique de synonymes* (cf. section 5) et au *filtrage sémantique* (cf. section 6).

2. Travaux antérieurs

Plusieurs méthodes ont été proposées ces dernières années pour mesurer la similarité entre des mots et développer des applications sémantiques dans le domaine du traitement automatique du langage. Elles peuvent être classifiées dans 4 catégories :

- celles qui exploitent un dictionnaire explicatif (Kozimo et Furugori, 1993 ; Kozima et Ito, 1995 ; …) ;
- celles qui exploitent Wordnet (Hisrt et St-Onge, 1997 ; Leacock et Chodorow, 1998 ;…) ;
- celles qui exploitent Wordnet et un corpus (Jiang et Conrath, 1997 ; Resnik, 1995 ; Lin, 1998c ;…) ;
- celles qui exploitent un thésaurus (Okumura et Honda, 1994 ;…).

Notre méthode se situe dans la première catégorie.

3. Similarité basée sur la Quantité d'Information Echangée (QIE)

La nouvelle définition de similarité que nous proposons se base sur des réseaux d'interconnexion entre des concepts et le contenu informationnel de ces concepts. Dans un réseau de ce type, les concepts jouent le rôle des nœuds et les rapports entre ces concepts sont représentés par les arcs. Nous expliquerons plus loin comment nous construisons un tel réseau. L'information contenue par les concepts et celle des rapports entre concepts seront consi-

dérées seulement au niveau quantitatif (i.e. la quantité d'information) mais pas au niveau qualitatif (i.e. la sémantique du type d'information). Ces idées initiales nous offrent une intuition très importante pour la définition de notre mesure de similarité.

Intuition : la similarité entre le concept A et le concept B dépend de l'information que A peut transférer vers B et de l'information que B peut transférer vers A. Autrement dit, la similarité de A et B dépend de la *Quantité d'Information Echangée (QIE)* entre A et B.

La supposition qui mène à une nouvelle définition de la similarité est la suivante :

Supposition : la description d'un concept A est constituée par la quantité d'information que ses objets voisins lui transfèrent.

Considérons un exemple avec un concept A qui est en rapport avec n autres concepts O_1, O_2, …, O_n.

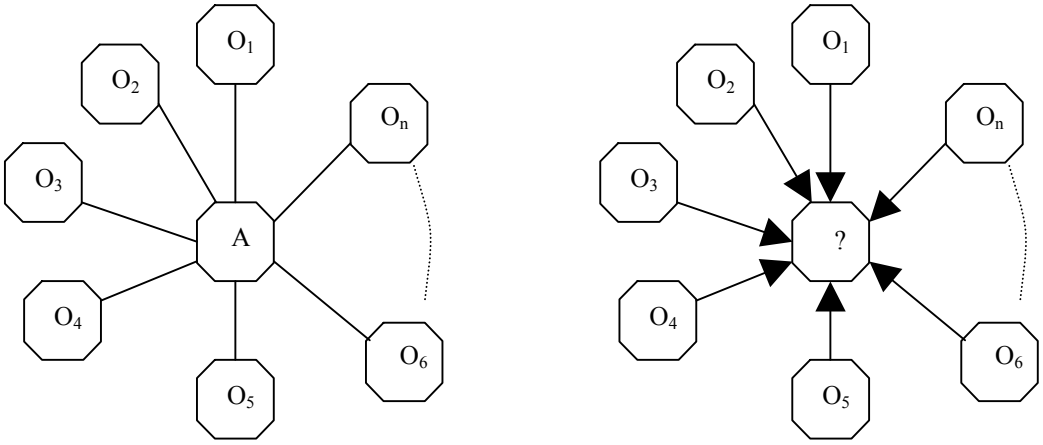

Figure 1. Un objet avec ses voisins

Dans la figure 1 à gauche, le concept A se trouve au milieu de ses voisins. Supposons que l'on ne connaisse pas A (cf. figure 1 à droite). Dans un tel cas, on ne connaît pas A, mais l'incertitude est réduite grâce à l'existence de O_1, O_2, …, O_n.

Et dans la théorie de l'information, l'information est toujours une mesure de la réduction de l'incertitude. Donc, on peut dire que les concepts O_1, O_2, …, O_n ont transféré une certaine quantité d'information pour réduire l'incertitude de A.

Dans un domaine où il existe un modèle probabiliste, et grâce à la théorie de l'information, le contenu d'information d'un concept A est calculé par la formule logarithmique de I :

$$I(A) = -\log(P(A)) \qquad (3.1)$$

où $P(A)$ est la probabilité de A. En partant de notre supposition, nous pouvons dire que la somme d'information que A reçoit de ses voisins est $I(A)$. On ne peut pas déterminer la quantité d'information qu'un O_i transfère vers A sur le lien de O_i à A, parce que cela dépend de la similarité entre des objets et que celle-ci n'a pas encore été définie. Mais on peut faire la remarque suivante :

Remarques : Un concept qui a plus d'information (faible probabilité d'occurrence) peut donner plus d'information.

Donc, on peut estimer la quantité d'information que chaque O_i peut transférer vers A sur le lien de O_i à A d'une manière très simple :

(Information transférée de O_i vers A) *(3.2)*

où $w_i = \frac{I(O_i)}{\sum_j I(O_j)}$

Maintenant, on va formaliser la similarité à partir des liens d'information que l'on vient de construire pour capturer notre intuition.

Considérons le cas où on a 2 concepts A et B dans un réseau d'interconnexion des concepts Ω.

- Le concept B peut transférer son information vers le concept A via un lien direct (B est un voisin de A) ou/et via des liens indirects (en passant par d'autres objets).
- Il existe au moins un chemin qui mène de A à B et vice versa.
- Les chemins qui connectent 2 concepts A et B décrivent la relation entre ces 2 concepts.

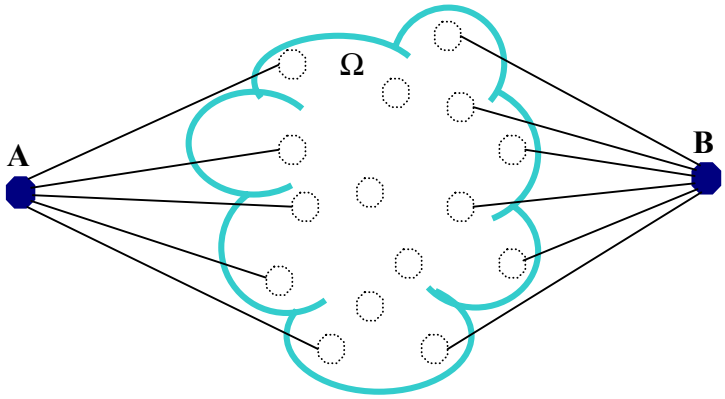

Figure 2. Deux concepts dans un réseau d'interconnexion

La quantité d'information que le concept B peut vraiment transférer vers le concept A via le réseau Ω est calculée par le *flot maximal d'information* de B à A, notée par $fmi_\Omega(B,A)$. Donc, la similarité entre A et B (qui correspond à la quantité d'information échangée entre A et B) est décrite par la formule :

$$sim_{QIE}(A,B) = f(mfi_\Omega(A,B), mfi_\Omega(B,A)) \quad (3.3)$$

Le choix du Ω dépend du domaine d'application. Les choix les plus naturels pour la fonction f sont la moyenne et la moyenne exponentielle. Et ça fournit les 2 formules suivantes :

$$sim_{QIE1}(A,B) = \frac{mfi_\Omega(A,B) + mfi_\Omega(B,A)}{2} \quad (3.4)$$

et

$$sim_{QIE2}(A,B) = \sqrt{mfi_\Omega(A,B).mfi_\Omega(B,A)} \quad (3.5)$$

Nous appelons cette mesure la similarité basée sur la *Quantité d'Information Echangée (QIE)*. Dans la section suivante, nous allons brièvement expliquer comment calculer la similarité entre des mots anglais. Ensuite, nous présenterons 2 applications de cette mesure de similarité : pour l'extraction de synonymes et pour le filtrage sémantique.

4. Similarité QIE pour les mots anglais

4.1. *Dictionnaire Webster pour l'anglais*

Grâce au projet Gutenberg, une version de dictionnaire 1913 US Webster (Webster) est publiée sur le Web dans 27 fichiers HTML. Chacun des 26 fichiers contient des mots commençant par le même caractère (A-Z), le dernier fichier ne contient que les mots nouveaux qui ne sont pas définis dans les 26 premiers fichiers.

Dans son rapport de stage à l'Université catholique de Louvain, P. Senellart (2001) décrit son travail de conversion du dictionnaire Webter en graphe et il analyse les propriétés de ce graphe. En lisant 27 fichiers HTML, il crée pour chaque mot du dictionnaire un nœud dans le graphe. Chaque arc (i,j) dans ce graphe représente l'occurrence de mot j dans la définition de mot i. Le graphe final comprend 112 169 nœuds et 1 398 424 arcs.

4.2. *Calcul de la similarité*

En ajoutant la capacité des liens au graphe Webster, nous avons un réseau de concept qui peut être utilisé pour calculer la similarité. En effet, nos applications utilisent rarement ce réseau complet pour calculer la similarité. Il suffit d'utiliser seulement un sous réseau qui se déduit avec tous les voisins de 2 mots dont nous devons calculer la similarité. Cette simplification nous permet d'accélérer la vitesse de calcul et d'éviter les connexions inutiles (trop longues connexions) entre les 2 mots.

4.3. *Évaluation de la similarité*

Comment peut-on juger si une mesure de similarité est performante et comment peut-on comparer deux mesures ? Ce sont des questions qui n'ont pas encore une réponse totalement convaincante.

La méthode qui nous semble la plus acceptable consiste à comparer les résultats avec les jugements humains. Dans le rapport Budanisky, 1999, 2 jeux de test ont été utilisés pour comparer plusieurs mesures de la similarité. Le premier jeu créé par Rubenstein et Goodenough (1965) contient 65 paires de mots. Le deuxième jeu créé par Millers et Charles (1991) contient 30 paires de mots. Toutes les paires dans le jeu de Millers et Charles se retrouvent dans le jeu de test de Rubenstein et Goodenough. Chaque paire de mots est associée avec un nombre qui indique la similarité jugée par des sujets humains. Pour évaluer les performances de la mesure computationnelle, le coefficient de corrélation entre les résultats de cette mesure et les jugements humains est calculé.

Le tableau 1 compare les résultats de notre méthode avec ceux d'autres méthodes. Les résultats numériques de ces mesures (Hirst-St-Onge, Jiang et Conrath, Leacock et Chodorow, Lin et Resnik) sont extraits de Budaniksy (1999).

Les résultats montrent que la méthode QIE peut donner des résultats équivalents aux autres méthodes en n'utilisant qu'un dictionnaire explicatif (en l'occurrence le Webster), tandis que d'autres méthodes utilisent des ressources plus structurées comme WordNet. Il faut mentionner en outre que nos résultats sont le fruit d'une première expérience et qu'ils pourraient être améliorés grâce à des traitements supplémentaires du dictionnaire Webster (reconnaissance des mots composés, lemmatisation, etc.). On peut donc considérer les résultats obtenus par notre mesure de similarité des mots comme prometteurs.

Méthodes	Rubenstein-Goodenough	Miller - Charles
Hirst et St-Onge	0.7861440344	0.7443990930
Jiang et Conrath	0.7812746298	0.8500267204
Leacock et Chodorow	0.8382296528	0.8157413049
Lin	0.8193023545	0.8291711020
Resnik	0.7786845861	0.7736382148
QIE1 (*f* = moyenne)	**0.7569047994**	**0.7515805974**
QIE2 (*f* = *moyenne exp.*)	**0.7859960606**	**0.8327221986**

Tableau 1. Coefficients de corrélation avec les jugements humains.

Dans les 2 sections qui suivent se trouvent deux applications qui utilisent cette similarité pour extraire des synonymes de mots anglais et pour filtrer les matériaux textuels.

5. Extraction de synonymes

L'extraction de synonymes est une application très utile. S'il est vrai que l'on peut facilement trouver des synonymes dans un thésaurus (comme *Roger*, par exemple) il n'en reste pas moins que de telles ressources n'existent pas dans toutes les langues, et que le nombre de synonymes proposés pour chaque mot est souvent limité. Par contre, un bon extracteur automatique de synonymes peut servir pour la plupart des langues et donner une longue liste de mots apparentés qui sont triés en ordre croissant ou décroissant de similarité.

La notion de « *synonymes* » dans ce sens concerne les mots liés entre eux et ayant un certain degré de similarité. Cette définition « large » convient également pour décrire les synonymes des thésaurus papier, car les synonymes que l'on y trouve ne peuvent pas être utilisés de manière interchangeable dans tous les cas.

Pour extraire les synonymes du mot *m*, nous calculons la similarité entre *m* et chaque mot dans le graphe des voisins de *m*. Ensuite, nous trions ces mots en fonction de ces similarités et proposons les *n* premiers mots comme synonymes de *m*.

Dans les tableaux 2 et 3, nous comparons les synonymes donnés par notre méthode avec ceux donnés par les méthodes présentées dans Senellart (2001) pour le mot *disappear* et le mot *sugar*.

Les synonymes de tous les mots du Webster ont déjà été extraits et peuvent être consultés grâce à une interface Web à l'adresse : http://cental.fltr.ucl.ac.be/synonyms.

6. Filtre sémantique

Le fonctionnement d'un filtre sémantique est caractérisé par l'élimination (ou la sélection) des éléments linguistiques (mot, phrase, texte, etc.) appartenant à une série de domaines thématiques spécifiques. Un domaine peut être décrit par une liste créé à la main de mots fortement liés à ce domaine (par exemple : finance={bilan, capital, capitaux, banque, bourse, compte, actions, opa, etc.}).

	Distance	**Senellart**	**ArcRank**	**QIE_L**	**WordNet**
1	Vanish	Vanish	Epidemic	Vanish	Vanish
2	Pass	Pass	Dissapearing	Fade	go away
3	Wear	Die	Port	Wear	End
4	Die	Wear	Dissipate	Die	Finish
5	Light	Faint	Cease	Pass	Terminate
6	Fade	Fade	Eat	Dissipate	Cease
7	Faint	Sail	Gradually	Faint	
8	Port	Light	Instrumental	Light	
9	Absorb	Dissipate	Darkness	Evanesce	
10	Dissipate	Cease	Efface	Disappearing	

Tableau 2. Synonymes de Disappear

	Distance	**Senellart**	**ArcRank**	**QIE_L**	**WordNet**
1	Cane	Cane	Granulation	Inversion	Sweetening
2	Starch	Starch	Shrub	Dextrose	Sweentener
3	Juice	Sucrose	Sucrose	Sucrose	Carbonhydrate
4	Obtained	Milk	Preserve	Lactose	Saccharide
5	Milk	Sweet	Honeyed	Cane	organic compound
6	Sucrose	Dextrose	Property	Sorghum	Saccarify
7	Molasses	Molasses	Sorghum	Candy	Sweeten
8	Sweet	Juice	Grocer	Grain	Dulcify
9	White	Glucose	Acetate	Root	Edulcorate
10	Plants	Lactose	Saccharine	Starch	Dulcorate

Tableau 3. Synonymes de Sugar

Le filtre le plus simple que l'on puisse imaginer est donc un programme qui éliminerait ou sélectionnerait des textes contenant au moins une occurrence de l'un des mots clés du domaine traité. Une approche naïve de ce type pose trois problèmes :

- Il est difficile d'établir ces listes de mots clés : comment faire, combien en faut-il ?
- L'absence de mot clé ne garantit pas qu'un texte n'appartient pas au domaine filtré.
- La présence d'un mot clé ne garantit pas que le texte appartient au domaine filtré (ambiguïtés, métaphores, etc.).

Nous avons essayé de trouver une solution qui exploite la notion de similarité des mots que nous venons d'exposer ci-dessus. Comment donc éliminer des textes appartenant à un domaine **D** en utilisant la notion de similarité des mots ? D'abord, pour chaque texte **A**, on va calculer la similarité entre chacun de ses mots et chacun des mots qui décrit le domaine **D**. Ensuite, la similarité entre le texte **A** et le domaine **D** est calculée en combinant toutes ces similarités. Cette valeur de similarité texte-domaine permet de créer les filtres plus flexibles que l'on peut paramétrer en déterminant un seuil d'acceptation (tous les textes qui reçoivent une valeur supérieure à ce seuil sont filtrés).

Une première version de filtre thématique a été expérimentée sur 1242 phrases extraites du journal américain en ligne *Detroit Free Press*. Nous avons défini le domaine à filtrer à l'aide des mots clés suivants : *crime, crimes, criminal, criminals, victim, victims, bail, bailed, bails, jail, jails, prison, prisons, custody, bribe, bribes, bribery, fraud, frauds, theft, forgery, drug, drugs, hashish, junkie, junkies, murder, murders, murderous, kill, killed, killing, killings, killer, hijacker, hijackers, hijack, hijacked, hijacking, kidnapping, kidnap, kidnapped, kidnapper, dying, homicide, homicides, suicide, suicides, terrorist, terrorists, hostages, hostage, prisoner, prisoners, genocide, genocides, atrocity, atrocities, brutal, vengeance, violent*[1].

À chacune des 1242 phrases est associée une valeur de « degré d'appartenance » qui est la somme des similarités (au sens défini ci-dessus) entre chacun des mot de la phrase et chacun des mots clés du domaine. Après avoir trié la liste des phrases en fonction de ce degré d'appartenance, nous pouvons constater que :

a) les phrases qui contiennent au moins d'un mot clé se situent principalement en tête de la liste et toutes figurent dans la première moitié.

Pos.	Appartenance	Mots clés	Phrase
1	139.852959	2	Two hours later, members of the city's Violent Crime Task Force saw the teenager and Denisha leaving the downtown Greyhound bus station with a man, Booth said.
2	118.085315	1	Yet the cost of housing, feeding and caring for a prison inmate is about $20,000 per year, or about $40 billion nationwide using 2002 figures, according to the Sentencing Project, a nonprofit organization in Washington, D.C., that promotes alternatives to prison.
3	115.218315	2	Since 1995, growth in the federal prison system mainly reflected more incarcerated drug offenders, accounting for nearly half of the total increase, and immigration offenders, accounting for more than 20 percent of the rise.
…	…	…	…

b) Parmi les premières phrases, on trouve des phrases qui ne contiennent pas de mot clé, mais dont le sens peut être effectivement considéré comme lié au domaine. Par exemple :

Pos.	Appartenance	Mots clés	Phrase
…	…	…	…
14	93.885475	0	One senator said 95 percent of the classified pages of a congressional report released last week into the work of intelligence agencies before the attacks of Sept.11, 2001, was kept secret only to keep from embarrassing a foreign government.
…	…	…	…
16	92.221736	0	ST.LOUIS (AP) -- A missing 23-month-old girl was found safe Sunday in Detroit with a teenage runaway more than a week after the

[1] Comme on le voit, il ne s'agit pas vraiment d'un « domaine », mais d'une liste assez hétéroclite de mots liés à des sujets ayant une connotation « violente » et étant fréquemment traités dans la presse.

			toddler was abducted from her St.Louis home, authorities said.
…	…	…	…
19	90.794144	0	In a hearing open to the public, government attorneys asked Haddad about Global Relief's links with Sheikh Abdallah Azzam, a man the government says cofounded Makhtab Al-Khidemat, a precursor to Al Qaeda, and was a mentor to Osama bin Laden.
…	…	…	…
26	84.262428	0	Hudson, 68, Gilbert, 55, and Platte, 66, were convicted in April of obstructing the national defense and damaging government property last fall after cutting a fence and walking onto a Minuteman III silo site in Colorado, swinging hammers and using their blood to paint a cross on the structure.
…	…	…	…

c) Les phrases qui se trouvent à la fin de la liste ne sont pas liées au domaine.

Pos.	**Appartenance**	**Mots clés**	**Phrase**
…	…	…	…
900	4.581682	0	Do your beard and your street rod suggest you've just left the set of a ZZ Top video?
901	4.533010	0	The group, with eight representatives and nine senators, has a Republican majority.
902	4.501355	0	Ultimately, Rochester Road also will be resurfaced between Gunn and Lakeville roads, but not widened.
903	4.476029	0	But they share space in a downtown Detroit office building and occasionally wind up on the same case.
…	…	…	…
1078	1.774052	0	"What's at stake is how good will the bill be?"
1079	1.767393	0	As well they might.
1080	1.751202	0	People are now safer, stronger
1081	1.720184	0	Night Pick 4 Numbers: two, eight, nine, one.
1082	1.711899	0	"Come and look at it," Josh said.
1083	1.710149	0	"But then again, we're excited to go see my husband.
1084	1.705655	0	Jacob Hoogendyk, R-Portage, was among lawmakers who said the subsidy should remain at $5.7 million.
1085	1.702224	0	Students at Livonia's Schoolcraft College will foot a 7 percent tuition increase.
…	…	…	…
1240	0.389159	0	I couldn't eat.
1241	0.387589	0	Haju Sunim Lundquist of the Zen Buddhist Temple in Ann Arbor.
1242	0.382479	0	There are lane closures on I-94 between Wayne and I-275.

Le programme fournit une liste ordonnée en fonction de la similarité par rapport au domaine filtré, ce qui représente un avantage par rapport au filtre naïf que nous avons évoqué au début

de cette section. La liste sera donc filtrée grâce à un seuil que l'on pourra déterminer en fonction des applications en jeu. Naturellement, plus ce seuil sera haut, plus grande sera la probabilité de sélectionner des éléments qui sont liés au domaine analysé.

Nous rendons compte dans cet article des premiers tests effectués avec ce filtre et les résultats ne sont que partiels, car les développements sont toujours en cours.

Bien entendu, en parcourant la liste, nous rencontrons également des exemples montrant que les performances sont encore loin d'être idéales. Le tableau suivant montre 2 exemples qui attestent des difficultés rencontrées.

Pos.	Appartenance	Mots clés	Phrase
…	…	…	…
4	112.147987	0	On Saturday, Terry Browning Jr., beat the previous top speed of 73.734 m.p.h by 0.364 m.p.h with the 1-liter modified boat the Legend from Virginia Beach, Va.. Tommy Thompson of Cambridge, Md, answered with 75.684 m.p.h on Sunday in For Sale.
…	…	…	…
909	4.316700	0	"If there had been any information they could connect him to terrorism, why would they remove him?"
…	…	…	…

Une des difficultés rencontrées pendant cet expérience est que le filtre est utilisé pour filtrer des phrases ayant une longueur très limitée (< 40 mots/phrase), donc, il n'y a parfois pas assez d'informations pour le filtrage. Les résultats devraient sans doute être meilleurs si le filtre était appliqué à des éléments linguistiques plus longs comme : des paragraphes, des articles, etc. En outre, plusieurs démarches devraient permettre d'améliorer les résultats :

- Utilisation d'un dictionnaire plus récent. Le Webster est un ancien dictionnaire dont le lexique est loin de refléter l'état du lexique de la presse d'aujourd'hui.
- Au lieu de la somme, trouver une meilleure combinaison des similarités entre les mots des phrases et du domaine.
- Tenir compte des mots composés.
- Amélioration de la mesure de similarité elle-même.

7. Conclusion

Dans cet article, nous avons introduit une nouvelle mesure de similarité basée sur la *Quantité d'Information Echangée (QIE)* dans des réseaux de concepts. La nouvelle mesure a été implémentée pour calculer la similarité entre mots (en anglais) en utilisant le dictionnaire *Webster 1913*. La comparaison numérique que nous avons réalisée montre que la précision de notre méthode est équivalente à celle d'autres méthodes existantes et même si ces méthodes utilisent de «meilleures» ressources comme par exemple WordNet. Par ailleurs, notre approche est facilement adaptable à la plupart des langues, car elle nécessite peu de ressources (un simple dictionnaire explicatif). Par conséquent, notre méthode peut être utilisée dans des applications qui doivent être portables dans plusieurs langues.

Notre extracteur de synonymes et le filtre sémantique sont des exemples concrets d'applications pouvant être développées sur ces bases. La qualité des résultats obtenus nous

permet de penser également à beaucoup d'autres applications comme : la recherche d'un mots à partir de son explication (un animal qui vole → c'est un oiseau), la mise en correspondance d'ontologies, etc.

Références

Budanisky A. (1999). Lexical Semantic Relatedness and Its Applications in Natural Language Processing. *Rapport Technique* CSRG-390, *Computer Research Group* – Université de Toronto.

Budanisky A. et Hirst G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second meeting of the North Americasimn Chapter of the Association for Computational Linguistics,* Pittsburgh.

Ford L.R. et Fulkerson D.R. (1962). *Flows in Networks*. Princeton Univ. Press.

Hirst G. et St-Onge D. (1997). Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum Chr. (Ed.), *WordNet: An electronic lexical database*. MIT Press.

Jannink J. et Wiederhold G. (1999).thesauruss Entry Extraction from an On-line Dictionary. In *Proceedings of Fusion '99*, Sunnyvale CA.

Jiang J. et Conrath D.W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th International Conference: Research on Computational Linguistics (ROCLING X)*, Academica Sinica : 19-33.

Kozima H. et Furugori T. (1993). Similarity between words computed by spreding activation on an English Dictionary. In *Proceedings of EACL-93 (Utrecht)* : 232-239.

Kozima H. et Ito A. (1995). Context-Sensitive Word Distance by Adaptive Scaling of a Semantic Space. In Mitkov R. et Nicolov (Eds), *Recent Advances in Natural Language Processing* (une serie de *"Contemporary Issues in Linguistic Theory" 136*). John Benjamins : 111-124.

Leacock Cl. et Chodorow M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In Fellbaum Chr. (Ed.). *WordNet: an electronic lexical database.* MIT Press : 265-283.

Lin D. (1998b). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING-ACL98*, Montréal.

Lin. D. (1998c). An Information-Theoretic Definition of Similarity. In *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin.

Miller G.A. et Charles W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes,* vol. (6/1) : 1-28.

Okumura M. et Honda T. (1994). Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion. In *Proceedings of Fifteenth International Conference on Computational Linguistics* (COLINGS-94), vol. (2) : 755-761.

Page L., Brin S., Motwani R. et Winograd T. (1998). The pagerank citation ranking : Bringing order to the Web. *Rapport Technique* Computer Science Department, Université de Stanford.

Resnik P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*.

Rubenstein H. et Goodenough J. (1965). Contextual correlates of synonymy. *CACM,* vol. (8/10) : 627--633.

Senellart P. (2001). *Extraction of information in large graphs – Automatic Search of Synonymes.* Rapport de stage. Université Catholique de Louvain.

Webster (2000). The Online Plain Text English Dictionary, http://msowww.anu.edu.au/~ralph/OPTED/, *dans le projet de Gutenberg*.