Conférence TALN 1999, Cargèse, 12-17 juillet 1999

Analyse et désambiguïsation lexicale par vecteurs sémantiques

Mathieu Lafourcade et Eugène Sandford

LIRMM – 161, rue Ada – 34392 Montpellier Cedex 5 - France
{lafourcade, sandford}@lirmm.fr
http://www.lirmm.fr/

Résumé

Cet article introduit une représentation du sens basée sur des vecteurs de notions. Ces vecteurs *sémantiques* ont pour but de rendre compte de l'ensemble des idées évoquées dans un segment textuel. Ce type de représentation utilisé en conjonction avec une analyse morphosyntaxique classique permet d'effectuer dans de nombreux cas une désambiguïsation lexicale efficace.

1. Introduction

Dans le cadre du traitement automatique de la langue naturelle (TALN), nous nous intéressons à la désambiguïsation lexicale sémantique couplée à une analyse morphosyntaxique (Rastier 1987). En effet, nombreuses sont les approches maintenant classiques en recherche d'information. Le traitement peut être statistique (Salton et MacGill 1983), dépourvu de syntaxe et à base de mot-clés (Salton 1988), basé sur différentes logiques (Sowa 1994), ou encore sur des taxonomies (Resnik 1995).

À partir d'un espace de notions élémentaires (issu d'un thésaurus), il est possible de construire des vecteurs (dit *sémantiques*) et de les associer à des mots. Les mots ambigus fusionnent les vecteurs correspondant aux différents sens. L'utilisation de vecteurs permet de se reposer sur des notions mathématiques classiques et il est ainsi possible d'effectuer des manipulations formellement bien définies auxquelles nous attachons des interprétations (linguistiques) raisonnables.

Nous définissons en premier lieu ce que sont les vecteurs sémantiques. Ces vecteurs sont attachés à des termes et sont directement issus d'informations que l'on trouve dans des thésauri. Afin d'affiner cette représentation, nous tenons compte des voisinages de notions correspondant à un ensemble de familles de sens. Cette opération de *compilation* d'ensembles de vecteurs sémantiques (dits *lexiques sémantiques*) est itérative. Ces vecteurs constituent ensuite la base de l'analyse sémantique dont le résultat (les vecteurs *recuits*) permet une désambiguïsation lexicale.

2. Généralités

La définition de *vecteurs sémantiques* a pour but d'obtenir d'une projection (sur un jeu de notions prédéterminé) des idées exprimées dans chaque segment textuel indépendamment du niveau syntaxique considéré (mots, syntagmes, phrases et textes).

Chaque composante d'un vecteur - une dimension de l'espace vectoriel - est associée à une notion issue d'un thesaurus. Pour le français, nous avons utilisé les 873 notions du thésaurus (Larousse 1992). L'objet n'est pas ici de discuter la pertinence du choix des concepts. Nous travaillons donc ici avec les vecteurs de dimension 873. Nous ne disposons pas d'une base

vectorielle de l'espace vectoriel engendré par les vecteurs sémantiques. Nous disposons par contre d'un ensemble générateur composé des vecteurs associé aux notions du thésaurus.

La description formelle des notions dans le thesaurus rentre dans le cadre de la théorie des champs sémantiques. Ces champs de sens se chevauchent sans qu'il soit possible pour l'instant de les différencier (Baylon et Mignot 1995) ou de donner une relation calculable entre ces champs. Par exemple, l'antonymie apparaissant entre les notions Accord et Désaccord du thesaurus pourrait supposer une opposition entre ces deux dimensions dans notre espace vectoriel. Leur coefficient serait donc respectivement à 1 et –1. Nous avons délibérément écarté cette approche car selon nous la négation d'un concept ne modifie pas la référence sémantique à celui-ci. L'antonymie reste une relation sémantique forte.

Le domaine de définition d'une dimension d'un vecteur sémantique est réel dans [0, 1]. Cette valeur représente l'intensité de la notion dans le vecteur sémantique (0 aucune activation, 1 activation maximale et unique). La norme des vecteurs est toujours égale à un entier fixé (mathématiquement 1).

3. Vecteurs sémantiques crus et cuits

Les composantes d'un vecteur cru sont directement issues du thesaurus ayant servis à l'indexation d'un terme et de son vecteur. Par exemple dans (Larousse 1992), le nom féminin *friction* se projette sur les notions suivantes : Frottement, Soins du corps, Désaccord, Conflit et de Coiffure (qui correspondent respectivement aux dimensions 228, 393, 429, 649, 867). Il se traduit sous forme de vecteur :

1 …	**228**		**393**		**429**		**649**		**867**	… 873
<0 … 0	$1/\sqrt{5}$	0 … 0	$1/\sqrt{5}$	0 … 0	$1/\sqrt{5}$	0 … 0	$1/\sqrt{5}$	0 … 0	$1/\sqrt{5}$	0 … 0>

Après une rapide expérimentation (sur un texte de mille mots), il nous est apparu que l'utilisation de vecteurs sémantiques bruts ne donnait pas de résultat particulièrement satisfaisant. Nous avons identifié les raisons suivantes :

- Alors qu'intuitivement deux notions ont clairement un rapport au moins indirect (par exemple Maladie et Médecine), ce rapport n'est pas explicite dans les vecteurs sémantiques crus. Certaines notions sont voisines dans une des hiérarchies d'un thesaurus, mais ce voisinage n'est pas exprimé dans le calcul des vecteurs sémantiques.
- La deuxième raison est le pendant de la première. Si certaines notions manquent, d'autres peuvent alors interférer pour choisir un sens inapproprié. Il nous faut donc rendre explicite leur voisinage pour diminuer l'intensité relative des notions non pertinentes.

Le voisinage permet d'augmenter un vecteur sémantique des notions proches de chaque notion présente dans ce vecteur. Nous appelons *cuisson* le processus d'augmentation d'un vecteur en fonction de son voisinage. Nous pouvons définir itérativement la fonction de cuisson comme suit :

$$V^{i+1}(x) = V^{i}(x) + \sum_{k=1}^{\dim(V)} \alpha^{i} \times V_k^{i}(x) \times V^{0}(\hat{tete}_k)$$

- $V^{i}(x)$ est le vecteur sémantique cuit à l'itération i. Le vecteur cru est noté $V^0(x)$.
- dim(V) est la dimension d'un vecteur sémantique. Ici, il s'agit d'une constante (de valeur 873 pour notre expérimentation).
- La notation $\hat{tete}_k$ correspond au mot représentant la notion d'indice k dans un vecteur V. Par exemple $\hat{tete}_{649}$ = « conflit »
- $V_k(x)$ correspond à la valeur de la k^{ieme} dimension de V(x).
- α^{i} correspond à un coefficient d'amortissement.

Cuire un vecteur V consiste à considérer tour à tour chaque dimension k, et pour chacune d'elle à additionner à V le vecteur cru correspondant au terme (le mot-tête) lié à la notion. L'intensité de ce vecteur est proportionnelle à la valeur de V sur la dimension k.

La cuisson d'un terme débute à partir de son vecteur cru (i = 0). On peut effectuer un nombre de *tours de cuisson* arbitraire. Le vecteur est normé entre chaque tour. Plus le nombre de tours est important, plus le vecteur cuit rendra fidèlement compte du voisinage.

En choisissant correctement la fonction alpha (décroissante, avec 0 comme limite), nous pouvons nous assurer d'avoir un processus de cuisson convergent (c'est-à-dire qu'au bout d'un nombre fini d'itérations, nous avons : $dist(V^{i+1}(x), V^{i}(x)) < \varepsilon$ aussi petit que l'on veut). *dist(V, V')* est une fonction de distance entre deux vecteurs *V* et *V'*.

Par exemple, pour le terme *frictionner*, nous avons (la norme des vecteurs est 1024) :

V^0 : frottement/512 propreté/512 nettoyage/512 coiffure/512

V^1 : nettoyage/511 propreté/510 coiffure/483 frottement/400 distinction/205 éviction/108 beauté/108 chirurgie/108 dessus/97 musique/76 poli/71 soins du corps/67 énergie/59 mécanique/59 son/46 élevage/41 description/37 conflit/36 désaccord/36 anormalité/36 bruit/32 …

V^2 : propreté/480 nettoyage/466 coiffure/421 distinction/329 frottement/306 beauté/179 chirurgie/175 éviction/158 dessus/149 musique/112 soins du corps/99 poli/89 énergie/82 son/80 mécanique/80 anormalité/72 élevage/58 sexe/58 description/53 conflit/53 désaccord/51 bruit/51 cri/49 pilosité/49 coup/45 production/43 enfance/43 blessure/42 jeu/40 supplice/37 …

4. Lexique sémantique

Le lexique sémantique est une collection d'association mots - vecteurs sémantiques. Un tel lexique provient d'un thésaurus et consiste à lister les notions directement attachées à un terme. La compilation du lexique consiste à cuire les vecteurs crus. Un lexique sémantique L contenant n termes permet trois opérations de base :

- trouver à partir d'un mot *w* (potentiellement augmenté d'informations morphosyntaxiques comme la catégorie, le genre, le nombre, etc.) son vecteur sémantique, si $w \in L$;
- l'opération inverse : à partir d'un vecteur sémantique V, trouver le mot *w* dont le vecteur sémantique est dans L le plus proche du vecteur V. C'est-à-dire w_i tel que :

$$w_i \in L \mid dist(V, V(w_i)) \leq dist(V, V(w_k)) \;\; \forall k = 1 \ldots n, \;\; w_k \in L$$

- à partir d'un vecteur sémantique, trouver l'ensemble de mots $\{ w_i \}$ tels que le vecteur sémantique soit suffisamment proche de V. C'est à dire $\{ w_i \}$ tel que :

$$\{w_i\} \subseteq L \mid dist(V, V(w_i)) \leq seuil$$

Les mots monosémiques sont associés à un seul vecteur sémantique. Les mots polysémiques sont liés à plusieurs sens dans le lexique sémantique et chacun de ceux-ci à un vecteur cru. Tous les sens sont énumérés sous le mot polysémique et le vecteur sémantique global du mot considéré est alors la somme (normée) des vecteurs sémantiques des sens. Par exemple (entre crochets, figurent comme commentaires les notions *en clair*) :

```
("friction" :N :FEM                        ("friction" :N :FEM  228 393 429 649 867
      ("frottement" 228)          →              ("frottement" 228)
      ("massage" 393 867)                        ("massage" 393 867)
      ("conflit" 429 649))                       ("conflit" 429 649))
```

À chacun des trois sens de *friction*, est associée une liste de numéro de notion. Pour l'analyse sémantique, nous regroupons (en conservant les éventuelles duplications) les notions associées aux sens des mots. C'est ce vecteur cru qui est ensuite cuit :

V^0 : frottement/447 soins du corps/447 désaccord/447 conflit/447 coiffure/447

V^1 : frottement/472 conflit/444 désaccord/365 soins du corps/349 coiffure/344 guerre/187 énergie/180 mécanique/177 musique/173 tribunal/167 inimitié/138 poli/127 adversité/116 dessus/101 machine/65 révolution/64 défiance/57 mouvement/54 conseil/50 guérison/46 divorce/45

V^2 : conflit/387 désaccord/312 guerre/296 frottement/278 soins du corps/258 tribunal/254 coiffure/249 énergie/221 mécanique/219 musique/214 inimitié/201 adversité/197 dessus/150 machine/138 révolution/134 mouvement/123 poli/118 politique/116 conseil/116 manœuvre/104 passe-temps/99 force/90 défiance/87 détection/82 enseignement/75 guérison/73 discordance/73 divorce/70 transport par route/69 son/68 médecine/66 dissemblance/64 paix/61 profane/60 jeu/53 maladie/51 vigueur/47 difficulté/46 main-d'œuvre/43 choc/43 repos/38 immunité/38 …

La sélection d'un vecteur à partir d'un terme peut être faite en tenant compte de la d'attribut (comme la catégorie morphosyntaxique, le genre, etc.) du terme. L'inverse est aussi vrai, car on souhaite aussi pouvoir trouver par exemple le substantif correspondant à un vecteur donné.

5. Vecteurs sémantiques recuits et analyse sémantique

Avant l'analyse sémantique, les nœuds de l'arbre de l'analyse morphosyntaxique correspondant aux termes du segment textuel sont affectés de vecteurs sémantiques cuits. L'analyse sémantique (la recuisson) consiste à effectuer des propagations et des fusions des différents vecteurs. Les vecteurs recuits ont alors remplacé les vecteurs cuit dans l'arbre d'analyse. Commençons par un exemple :

(1) Après la publication *des notes d'examen*, on a pu constater une grande effervescence chez les étudiants.

En sortie de notre système, le syntagme prépositionnel (2) « *des notes d'examen* » a le vecteur sémantique suivant :

enseignement/334 parole/191 discours/176 jeu/174 recherche/168 informatique/160 philosophie/158 théologie/152 communication/151 médecine/135 musique/127 échec/125 son/118 grammaire/108 apprentissage/105 tentative/99 religion/99 morale/98 loyauté/98 détection/98 préparation/95 dieu/93 discordance/93 passe-temps/92 contrôle/92 musicien/87 commencement/86 livre/85 cause/84 question/83 emploi/82 enfance/82 agitation/81 langue/80 passion/80 mot/78 rapidité/78 presse/77 théâtre/76 regret/76 mouvement/76 habitude/74 supplice/72 multitude/72 conseil/71 condamnation/70 voyage/68 bulle/68 …

La structure syntaxique du texte ordonne par niveaux les points suivants : le texte correspond à la racine de la structure syntaxique. Elle a comme fils, les points représentant les phrases des textes. Les points correspondant aux syntagmes se trouvent sous les points des phrases. Au niveau le plus bas de la structure, se trouvent alors les points – les feuilles – correspondant aux mots du texte. Afin d'expérimenter plusieurs approches pour l'analyse sémantique, nous avons défini plusieurs fonctions Fi de propagation.

- La fonction *F1*, consiste à propager les vecteurs sémantiques des mots vers la racine de la structure pour pouvoir calculer l'ensemble du contexte sémantique du texte puis de raffiner les vecteurs sémantiques des mots en fonction de ce contexte.
- La fonction *F2* possible consiste à calculer le vecteur sémantique du mot au fur et à mesure de la construction de gauche à droite. Elle correspond à une construction séquentielle du contexte.
- La fonction *F3* calcule les vecteurs sémantiques des syntagmes puis des phrases et du texte. Mais au lieu d'effectuer une comparaison directe des vecteurs sémantiques des mots avec les phrases et/ou le texte comme en *F1*, elle redescend vers le bas de la

structure en affinant les vecteurs sémantiques des syntagmes et enfin, les vecteurs sémantiques des mots. Cette opération peut être itérée.

Dans les expérimentations actuelles, nous utilisons une combinaison de ces fonctions. Chaque point de la structure s'enrichit de la représentation sémantique vectorielle des points dont il est voisin structurellement (ascendant, descendant, frère, etc.). Les segments en incise ne perturbent donc pas les calculs de voisinage.

L'augmentation sémantique de la structure syntaxique se fait d'abord du bas vers le haut, éventuellement de façon transversale (gauche-droite et droite-gauche) et enfin en redescente. La remontée des notions vers le haut de la structure s'effectue par une fusion des notions dans les vecteurs. Cette fusion est généralement une combinaison linéaire. Cette sommation dépend des fonctions syntaxiques des mots dans l'arborescence. Le vecteur obtenu à la racine de la structure est alors le vecteur représentant le contexte sémantique de l'ensemble du texte.

Deux choses sont à considérer à ce niveau : le point peut-il intégrer ou non le vecteur sémantique et si oui, de quelle manière ? Le vecteur sémantique importé du lexique sémantique, est intégrable au point considéré si tous les descendants du même point sont marqués. Si elle est possible, l'intégration revient à une combinaison linéaire des vecteurs sémantiques des points-fils plus le résultat de la lecture du lexique sémantique. Cette lecture se justifie par le fait qu'un lexème complexe non figé peut être indexé dans le lexique. Ce résultat est soit (1) nul si aucun schéma n'est reconnu appartenant au lexique sémantique, soit (2) le vecteur sémantique si le schéma est reconnu. Si un élément non marqué *x* a son vecteur sémantique V intégrable alors :

$$V = Lecture(x, L) + \sum_{y \in D} k \times V(y) \qquad \text{où :}$$

- *Lecture* est une fonction qui nous renvoie le vecteur sémantique issu du lexique sémantique L devant être intégré au point *x* de l'arborescence considérée.
- D = Descendants(x) ;
- *k = 1* si la fonction syntaxique du point associé au vecteur est gouverneur du groupe considéré, sinon k<1 (en général 1/2). Le vecteur sémantique V représentant le groupe les contenant, doit donc tenir compte de cette particularité.

La structure syntaxique *se sémantise* selon la propagation des vecteurs et constitue le processus d'augmentation lexicale sémantique.

6. Désambiguïsation sémantique

La recherche du sens d'un mot polysémique est la sélection des sens indexés dans le lexique sémantique. L'idée est de comparer chaque sens indexé avec le vecteur recuit et aussi les contextes de la phrase (contexte local) ou du texte (contexte global). Ce sont les vecteurs sémantiques associés aux points représentant le texte et la phrase. Nous souhaitons que cette comparaison soit exprimée par une distance mathématique dans notre espace vectoriel *E*. Outre la distance euclidienne (dont l'interprétation est difficile), nous pouvons aussi calculer la distance en termes de longueurs de chemins dans la hiérarchie du thesaurus (dont l'interprétation est encore plus douteuse et surtout dépend de l'ontologie).

Une bonne fonction de distance doit être liée à la proximité sémantique. Celle-ci peut être raisonnablement représentée par le cosinus de l'angle de deux vecteurs. Plus deux entités – mot, syntagme, phrase ou texte – évoquent les mêmes idées, plus leurs vecteurs associés sont colinéaires (c'est-à-dire que $\forall V_1, V_2 \in E, \cos(V_1, V_2) \to 1$). Or le *cosinus* dans notre espace vectoriel n'est pas une distance au sens mathématique (car *cosinus*(x, x) $\neq 0$).

Tous les vecteurs de notre espace E sont normés et leurs coefficients sont positifs ou nuls. Par conséquent, l'angle que forment deux vecteurs sémantiques quelconques est compris entre 0 et $\pi/2$. La fonction *ArcCosinus* constitue alors une distance au sens mathématique dans *E*. La fonction de distance *dist* que nous utilisons est donc définie ainsi :

$$\forall V_1,\ V_2 \in E,\ \text{dist}(V_1,\ V_2) = \text{ArcCosinus}(V_1,\ V_2)$$

Sélectionner le sens le plus proche d'un vecteur donné consiste alors à comparer ce vecteur à ceux du lexique sémantique et à retenir celui dont la distance est minimale.

7. Conclusion

L'approche proposée ici est en cours d'évaluation et les lexiques sémantiques sont indexés conjointement. Les premiers résultats sont particulièrement encourageants. Nous avons testé notre approche sur un corpus très réduit de phrases contenant des termes typiquement ambigus (du type de celles présentées dans cet article). Dans l'état actuel de l'indexation, nous arrivons à sélectionner le sens et le terme correct dans 2/3 des cas. Auparavant, la sélection se serait faite au hasard. Les cas typiques d'échec sont ceux où ce le contexte lexical ne suffit pas. Dans certains cas (rares), il peut aussi renforcer l'erreur. Un exemple typique serait « L'intérêt qu'il retire de son étude sur les banques est évident ». Pour le terme « intérêt », le sens de « profit/somme d'argent » se trouve particulièrement renforcé à cause de la présence de « banques ». Nous comptons poursuivre et affiner notre étude. D'une part, l'indexation du lexique sémantique du français et de l'anglais doit se poursuivre.

Nous avons présenté dans cet article, une approche de TALN basée sur des vecteurs sémantiques. De tels vecteurs représentent avec finesse l'ensemble des idées évoquées dans un segment textuel. Il est ainsi souvent possible, en cas d'ambiguïté lexicale de sélectionner le sens correct d'un terme.

Références

BAYLON C. et MIGNOT X. (1995) Sémantique du langage : initiation. Editions Nathan, Paris.

CHAUCHE J. (1984) Un outil multidimensionnel de l'analyse du discours. Dans Proceedings of the International Conference COLING'84, Standford, July 1984.

KIRKPATRICK B. (1986) Roget's Thesaurus of English Words and Phrases. Penguin Books.

LAROUSSE (1992) Thésaurus Larousse – des idées aux mots – des mots au idées. Larousse, ISBN 2-03-320-148-1.

RASTIER, F. (1987) Langages : Sémantique et Intelligence Artificielle. Bernard Willerval Jouve, Septembre 1987.

RESNIK P. (1995) Using Information content to evaluate semantic similarity in a taxonomy. Dans Proceedings of IJCAI-95.

SANDFORD E. et FRAÏSSE S. (1995) Un outil d'extraction de la sémantique d'un corpus textuel Dans actes des IVèmes journées scientifiques de l'UREF, Lyon, 28-30 septembre 1995.

SALTON G. (1988) Term-Weighting Approaches in Automatic Text Retrieval. McGraw-Hill computer science serie. McGraw-Hill, Volume 24.

SALTON G. et M. J. MACGRILL (1983) Introduction to Modern Information Retrieval. McGraw-Hill computer science serie. McGraw-Hill, New-York.

SOWA J. F. (1984) Conceptual Structure. Addison Wesley.

SPARCK JONES K. (1986) Synonymy and Semantic Classification. Edinburgh Information Technology Serie.