RECITAL 2003, Batz-sur-Mer, 11–14 juin 2003

Annotation sémantique hors-source à l'aide de vecteurs conceptuels

Fabien JALABERT
LIRMM (CNRS - Université Montpellier 2)
Laboratoire d'Informatique, de Robotique
et de Microélectronique de Montpellier
161, rue Ada - F - 34392 Montpellier Cedex 5
jalabert@lirmm.fr

Mots-clefs – Keywords

annotation sémantique, désambiguisation sémantique lexicale
WSD , word sens disambiguation, word sens tagging, annotation

Résumé - Abstract

Dans le cadre de la recherche en sémantique lexicale, nous utilisons le modèle des vecteurs conceptuels pour représenter les sens de termes. La base vectorielle est construite à partir de définitions provenant de diverses sources lexicales, ce qui permet statistiquement de tempérer les diverses incohérences locales. Pour désigner le sens obtenu après un regroupement des définitions, nous utilisons un identificateur qui entraîne certaines contraintes. En particulier, un "cluster" de définition est désigné par une référence vers différentes définitions de la multisource. D'autre part, le contrôle de la qualité d'une classification ou désambiguisation de sens impose de faire référence en permanence au lexique source. Nous proposons donc de nommer un sens à l'aide d'un autre terme du lexique. L'annotation est un outil léger et efficace qui est essentiellement une association d'idées que l'on peut extraire de toute base de connaissance linguistique. Les annotations obtenues peuvent finalement constituer une nouvelle source d'apprentissage pour la base de vecteurs conceptuels.

In the framework of research in meaning representation in NLP, we focus our attention on thematic aspects and conceptual vectors. This vectorial base is built by a morphosyntaxic analysis of several lexical resources to reduce isolated problems. Also a meaning is a cluster of definitions that are pointed by an Id number. To check the results of an automatic clustering or WSD, we must refer continously to the source dictionnary. We describe in this article a method for naming a word sens by a term of vocabulary. This kind of annotation is a light and efficient method the uses meanings associations someone or something can extract from any lexical knowledge base. Finally, the annotations should become a new lexical learning resource to improve the vectorial base.

Introduction

Dans le cadre de la recherche en sémantique lexicale, l'équipe TAL du LIRMM développe actuellement un système d'analyse des aspects thématiques des textes et de désambiguisation lexicale basé sur les vecteurs conceptuels. Les vecteurs représentent les idées associées à tout segment textuel (mots, expressions, textes, ...) via l'activation de concepts. Pour la construction des vecteurs, nous avons pris deux hypothèses principales : l'*automatisation* de la création de la base lexicale vectorielle par apprentissage à partir d'informations extraites de diverses sources (dictionnaires à usage humain, liste de synonymes, ...), et un *apprentissage multisource* afin de palier au bruit définitoire (par exemple les problèmes dûs au métalangage comme dans la définition d'‹*aboyer*›, *crier en parlant du chien*). Chaque dictionnnaire proposant un découpage des sens pour une entrée, nous avons mis en place une procédure qui associe à un sens un ensemble définitions. Ce groupe de définitions est désigné par un identifiant numérique utilisé lors d'un processus de désambiguïsation. L'utilisateur, humain ou machine, doit donc connaître le codage pour pouvoir réassocier à un terme et un identifiant le sens correspondant. Chaque utilisateur doit dès lors posséder les sources lexicales du ou des désambiguïseurs auxquels ils font appel. Superviser une désambiguïsation manuellement demande systématiquement de consulter la source pour chaque terme polysémique. Nous proposons dans cet article de nommer un sens par un terme de la langue. Tout agent possédant une compétence linguistique doit être capable de retrouver le sens uniquement à partir du terme annoté (couple (terme, annotation)). Une telle procédure offre l'avantage d'une bonne interopérabilité, différents désambiguïseur peuvent proposer leurs résultats à différents clients (traducteur, indexeur, ...) et ces derniers peuvent faire appel à de multiples désambiguïseurs pour pallier aux lacunes de certains. Nous présentons, dans un premier temps, une formalisation de l'annotation puis le modèle des vecteurs conceptuels. Nous détaillerons ensuite la procédure d'annotation que nous proposons et qui repose à la fois sur les vecteurs conceptuels et sur les sources lexicales.

1 Annotation sémantique, nommage de sens : définition

Les systèmes actuels associent à chaque mot polysémique une annotation numériques, par exemple : "*Le chat*/**I.2**/ *mange*/**1**/ *la souris*/**II.1**/". L'usage que nous proposons diffère des précédents, nous souhaitons retrouver des associations d'idées dans la langue et les utiliser pour désigner un sens. Nous associons à chaque terme polysémique d'un texte un annoteur qui est lui même terme du lexique. Par exemple :

Chaussé/**porter**/ de ses *bottes*/**chaussure**/ il *revenait*/**déplacer**/ vers la grange et *apercevait*/**voir**/ les *bottes*/**amas**/ de *foin*/**paille**/.

Le lexicographe qui possède ainsi un annotateur et un terme sera alors capable d'identifier plus facilement le sens désigné. Il ne s'agit plus de donner une définition comme annotateur (Wilks, Stevenson, 1997) mais d'extraire un représentant. Par exemple, nous proposons d'annoter le terme ‹*botte*› par *botte*/**paille**/, *botte*/**chaussure**/, *botte*/**escrime**/ qui sont des formes plus intuitives et compréhensibles. On peut alors considérer que cette annotation est *hors-source*, elle fait sens sans source lexicale spécifique. Si le destinataire a une compétence linguistique, il lui sera possible de réassocier la botte et la chaussure, ou la botte et la réunion de végétaux.

1.1 Définition formelle

L'annotation correspond à une fonction bijective qui, à un terme et un sens, associe un couple *(terme, annotation)*. Soit D le dictionnaire, M un terme du dictionnaire, s_i un sens de M tel que $s_i = M$ et A_i un ensemble d'annotateurs pour s_i et soit f la fonction d'annotation :

$$f : \forall M \in D, \forall s_i \in M \qquad s_i \longrightarrow A_i \qquad A_i \subset D$$
$$\text{telle que : } \forall A_i = f(s_i), \quad \forall\, A_j = f(s_j), \quad i \neq j, \quad A_i \cap A_j = \emptyset \quad et \quad A_i, A_j \neq \emptyset$$

Quel que soit le terme $M \in D$, on partitionne le lexique pour obtenir pour chaque sens $s_i \in M$ un ensemble (non vide) d'annotateurs A_i. Cette fonction doit permettre de réassocier à tout couple (a, M) où a est un élément de A_i le sens s_i correspondant. Cette fonction admet donc une bijection réciproque f^{-1} :

$$f^{-1} : \forall M \in D, \qquad \forall A_i \subset D, \qquad \forall a \in A_i \qquad (a, M) \longrightarrow s_i$$

On souhaite ainsi insérer dans un texte des balises contenant un terme discréminant pour chaque terme polysémique désambiguïsé. La fonction d'annotation précédente propose pour un sens donné d'un terme un ensemble de candidats. Il importe donc d'évaluer les différents candidats et de les classer par ordre d'intéret tout en accordant une souplesse suivant l'utilisation qui sera faite de l'annotation (usage humain, interopérabilité, ...).

1.2 Propriétés remarquables

1.2.1 Indépendance aux dictionnaires

Dans le cas où on ne souhaite pas diffuser le dictionnaire source (droits d'auteur, volume trop important,...) une annotation par des termes de la langue peut permettre au client de retrouver le bon sens sans la source. Si dans plusieurs dictionnaires un terme est associé à un autre, il est fortement probable de retrouver la même relation à l'aide d'autre sources. Différents annotateurs n'offriront pas la même indépendance aux sources, il est donc important d'évaluer cette propriété et de classer les annotateurs en conséquence.

1.2.2 Une interface homme-machine

L'annotation est un outil précieux pour le lexicographe qui supervise une désambigïsation. Elle met en évidence de façon simple le sens qu'il faut attribuer à un terme, sans devoir constament se référer à un dictionnaire donné et en assimiler la définition. Le coût cognitif est défini comme l'effort que doit faire l'agent humain pour transformer une perception, une information en une connaissance exploitable (Prince, 1996). Dans le cas de l'annotation sémantique à usage humain, il est déterminant de minimiser ce coût et pour cela de prendre en compte d'autre critères. On favorisera ainsi des candidats qui ont un usage proche en utilisant des notions de fréquence, de cooccurence terme/annotation, mais aussi des informations fournies par les dictionnaires comme la morphologie ou l'usage (contexte ou domaine, sens figuré ou soutenu, ...).

2 Les vecteurs conceptuels

Le modèle vectoriel a été introduit par (Salton, 1968) en recherche d'information. C'est à partir de (Chauché, 1990) que l'on a une formalisation de la projection de la notion linguistique de champs sémantiques dans un espace vectoriel. A partir d'un ensemble de notions élémentaires dont nous faisons l'hypothèse, les concepts, il est possibles de construire des vecteurs (dits conceptuels) et de les associer à des items lexicaux. Dans notre expérimentation sur le Français nous utilisons (Thésaurus Larousse, 1992) dans lequel sont définis 873 concepts. L'hypothèse générale du thésaurus, que nous adoptons ici, est que cet ensemble constitue un espace générateur (non libre) pour les termes et leur sens. On définit *Une mesure de similarité* $Sim(X, Y) =$

$\cos(\widehat{X,Y}) = \frac{X \cdot Y}{\|X\| \times \|Y\|}$ avec "$\cdot$" désignant le produit scalaire et la *distance angulaire* $d_A(X,Y) = \arccos(Sim(X,Y))$. Cette dernière mesure représente intuitivement une notion de distance thématique entre deux mots, elle respecte la symétrie, la réflexivité et l'inégalité triangulaire. Voici un exemple avec le terme *amour* où les termes les plus proches sont *éprendre* (0.17), *Cupidon* (0.25), *Aphrodite* (0.27), *amitié* (0.308), *tendresse* (0.3084). Les concepts les plus activés sont *amour* (11879), *passion* (4137), *inimitié* (3511), *courtoisie* (3348),... La construction de la base vectorielle est automatisée à partir de définitions issues de dictionnaires à usage humain (Schwab & al., 2002), (Lafourcade & al., 2002).

3 La procédure d'annotation

Notre procédure d'annotation repose sur une utilisation mixte des ressources lexicales et des vecteurs conceptuels. Nous l'avons vu, nous souhaitons annoter avec un terme du lexique. Pour des raisons pratiques (coût calculatoire élevé), nous utilisons dans un premier temps les sources lexicales pour extraire les candidats. Une procédure de validation permet de retirer tous ceux qui ne satisfont pas aux conditions de bijectivité (cf. 1.1). Enfin, pour obtenir une pertinence et un coût cognitif optimaux, des évaluateurs ordonnent le résultat suivant divers critères (fréquence, cooccurence, morphologie, ...).

3.1 Extraction de candidats

On distingue plusieurs types d'extracteurs qui sont caractérisés par leur source et la méthode employée. Les premiers s'appliquent à un ensemble de dictionnaires traditionnels, analysant les différentes définitions et proposant tous les mots contenus comme candidats. Nous utilisons pour cela SYGMART [1], un analyseur morphosyntaxique calculé à partir d'une phrase donnée un arbre morphosyntaxique donnée. On obtient alors des informations complête sur la morphologies des termes de la phrase ainsi que leur fonction. On filtre articles, pronoms, métalangage (*famille de, du latin,...*) qui n'apportent pas d'information sémantique à la définition. D'autre extracteurs ont une méthode identique mais s'appliquent à des sources comme une liste de synonymes (Ploux, Victorri, 1998) , concepts du thésaurus, de dictionnaires de cooccurence,... Enfin, les derniers extracteurs s'appliquent à des dictionnaires dont on a pu retirer une information structurée grâce à une analyse morphosyntaxique ou à un format semi-structuré comme XML. On peut alors retrouver des relations comme synonymie, antonymie, hyperonymie, règles d'usage, exemples caractéristiques,... Dans le cas des noms propres, on détermine qu'il s'agit d'une ville, d'un fleuve d'un pays précis, etc. Ces extracteurs travailleront ainsi sur tous les cas particuliers qu'on peut extraire des définitions et s'avèrent particulièrement efficaces sur les noms propres.

3.2 Validation d'un candidat

Après avoir extrait de nombreux candidats, il est nécessaire de les évaluer. Cependant, avant de classer des candidats, un filtre va valider ou non des candidats. Comme nous l'avons expliqué précédemment, nous souhaitons obtenir une fonction admettant une réciproque pour associer à un couple (a_i, M) un sens $s_i \in M$. L'objectif de la procédure de validation est simplement d'ôter tout candidat qui ne satisfait pas cette condition. Elle se formalise de la façon suivante : un annotateur a est valide si et seulement si:

$$\forall s_j \neq s_i, \qquad D_A(a, s_i) \leq D_A(a, s_j)$$

[1] développé par Jacques Chauché : *http://www.lirmm.fr/∼chauche/Pr%E9sentationSygmart.html*

3.3 Evaluation d'un candidat

3.3.1 Une note d'extraction des candidats

Durant la phase d'extraction des candidats, il est déjà indispensable de les évaluer suivant : (1) *la source lexicale* (différents dictionnaires n'offrent pas la même qualité de résultat), (2) *le type d'extraction* (les concepts les plus activés ou les domaines et usages (qui ne sont pas le plus souvant les meilleurs) et enfin (3) *l'arbre morphosyntaxique* fournit des informations comme la position et la fonction du terme dans la phrase qui sont particulièrement précieuses. En effet, les premiers termes d'une définition contiennent très souvent un excellent annotateur du fait qu'elles sont exprimées en genre et différence. SYGMART nous permet de tenir compte particulièrement du sujet ou du complément d'objet dans la définition et de résoudre les difficultés engendrées par l'apposition de compléments.

3.3.2 La distance angulaire et l'annotation

Pour assurer la reciprocité de l'annotation, il est indispensable que le candidat soit le plus proche possible du sens annoté tout en maximisant sa distance avec tout autre sens. La notion de distance angulaire permet d'évaluer les candidats potentiels grâce à trois mesures :

La marge de désambiguïsation absolue. Elle représente l'intervalle dans lequel la fonction réciproque n'associe pas l'annotateur à un mauvais sens. Plus cette marge est importante, plus les probabilités sont bonnes de ne pas réassocier un mauvais sens dans une autre base lexicale. Soit a_1 l'annotateur de s_1 et $s_2 \neq s_1$ le second sens le plus proche de a_1. Alors la marge absolue est :

$$MARGE_A(a_1, s_1) = \mid D_A(a_1, s_2) - D_A(a_1, s_1) \mid$$

La marge de désambiguïsation relative. Prenons comme exemple deux candidats a_1 et a_2 qui annotent un sens $s \in M$ avec les valeurs suivantes : $D_A(a_1, s_1) = 0.15$, $D_A(a_1, s_2) = 0.33$, $D_A(a_2, s_1) = 0.30$ et $D_A(a_2, s_3) = 0.50$. Leur marge de désambiguisation respectives sont : $MARGE_A(a_1, s_1) = 0.18$ et $MARGE_A(a_2, s) = 0.20$. La marge absolue favorise ainsi le deuxième candidat. Pourtant, le premier est bien mieux associé au sens annoté. La marge relative prend en compte la distance entre le candidat est le sens nommé de la façon suivante : $MARGE_R(a_1, s) = \frac{MARGE_A(a_1,s)}{d_1}$

Le risque de non-sens : Cette dernière mesure possède encore un dernier défaut : elle ne tient pas compte de la distance entre les deux sens les plus ambigus. Prenons l'exemple de ‹*frégate*›. Les trois sens de ce mot sont (1) ‹*Oiseau de mer*›, (2) ‹*bâtiment de guerre à trois mâts*› et (3) ‹*Bâtiment d'escorte anti-sous-marin*›. Soient les deux candidats ‹*voilier*› et ‹*guerre*› possédant une marge de désambiguisation identique qui annotent (2). ‹*Voilier*› a pour sens le plus proche (1) tandis que ‹*guerre*› l'est avec (3). Alors on doit tenir compte de la distance entre les sens ambigus. Ainsi la distance entre (1) et (2) étant bien plus importante qu'entre (2) et (3), une erreur que ferait la procédure réciproque serait bien plus importante si elle associait (1) à (2) dans un texte que si elle associait (3). Cependant, on peut souhaiter dans le contexte où les sens (2) et (3) seraient très présents et le sens (1) absent utiliser l'annotation (1), l'objectif étant de discréminer au mieux les sens (2) et (3).

3.3.3 Autres évaluateurs d'usage

Pour selectionner un terme réellement associé, il est déterminant de tenir compte de l'usage de l'annotateur et de l'annoté. Pour cela, plusieurs évaluateurs sont présents dans notre procédure :

(1) *La fréquence d'usage de l'annotateur* : supposons que ‹*mouche*› reçoive comme candidat *mouche*/**drosophyle**/ (pour le sens de l'insecte). Un annotateur très rare risque de ne pas être

connu du client, qu'il soit humain ou automatique, un annotateur trop fréquent risque de ne pas faire sens (*cuisine*/**faire**/). De même, suivant l'utilisation, on peut souhaiter dans l'ordre de préférence un hyperonyme, un hyponyme ou un co-hyponyme. La fréquence est un indice : un terme bien moins fréquent qu'un autre a plus de chance d'être un hyponyme qu'un hyperonyme.

(2) *La catégorie grammaticale* : le choix d'utiliser des termes de même nature grammaticale réduit le coût cognitif, technique qu'utilisent depuis longtemps les dictionnaires pour définir un terme (‹*déplanter*› → ‹*enlever*› ; ‹*rimer*› → ‹*constituer une rime*› ; ‹*table*› → ‹*meuble*› ...).

(3) *L'usage* : deux utilisateurs différents n'auront pas les mêmes associations d'idées entre différents mots. Par exemple, le terme ‹*police*› peut signifier le contrat d'assurance ou l'autorité judiciaire. Dans ce deuxième sens, l'annotation ‹*agent*› ou ‹*poulet*› n'induira pas le même comportement chez le lecteur. Celui-ci risque de ne pas retrouver une association pourtant évidente chez un autre. La cooccurence permet de confirmer l'association d'idée entre deux termes. De plus, un étude en contexte et en usage offrira de meilleurs résultats. Un site personnel et un livre ne s'adresseront pas aux même lecteurs, tout comme un article de presse et un article scientifique.

Conclusion et perspectives

L'annotation telle que nous venons de la présenter est un outil important dans le cadre de nos recherche. C'est une interface homme-machine permettant une évaluation rapide et efficace pour le superviseur d'un processus de désambiguïsation dans le cadre de la traduction, ou dans notre cas, une aide de l'analyse de définitions pour produire un vecteur conceptuel. Elle est, d'autre part, un nouvelle forme d'interfaçage entres de multiples systèmes. Les résultats actuels sont encourrageants, mais de nombreuses voies restent à explorer. Nous projetons ainsi dans un avenir proche d'étudier précisément les comportement de l'utilisateurs et d'expérimenter l'interfaçage entre différents systèmes automatiques dans un cadre multilingue. Enfin, cette analyse des associations d'idées va devenir une nouvelle source lexicale pour la base vectorielle. L'objectif est par cela d'améliorer la base de connaissance, et de ce fait notre propre processus (phénomène de la double boucle (Schwab, 2003).

Je remercie D. Schwab, M. Lafourcade et V. Prince pour l'aide précieuse qu'ils m'ont apportée

Références

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

Hachette. *Dictionnaire Hachette Encyclopédique.* Hachette, ISBN 2.01.280477.2, version en ligne: http://www.encyclopedie-hachette.com

M. Lafourcade, V. Prince, D. Schwab *Vecteurs conceptuels et structuration émergente de terminologies* Revue TAL Volume 43 - n 1/2002, pages 43 à 72

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

S. Ploux, B. Victorri *Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes* Revue TAL Volume 39, 1998, Numéro 1.

V. Prince *Vers une informatique cognitive dans les organisations - Le rôle central du langage* Ed. Masson 1996.

G. Salton *Automatic Information Organisation and Retrieval* McGraw-Hill, New York 1968.

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

D. Schwab *Société d'agents apprenants et sémantique lexicale : comment construire des vecteurs conceptuels à l'aide de la double boucle.* RECITAL 2003, Batz-sur-Mer, Juin 2003.

Y. Wilks, M. Stevenson *Sense tagging : semantic tagging with a lexicon* Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?, Washington, D.C. (1997).