Acquisition sur corpus non spécialisés de classes sémantiques thématisées

Mathias Rossignol[1], Pascale Sébillot[2]

mathias.rossignol@gmail.com, Pascale.Sebillot@irisa.fr

[1]MICA - 1 Đai Cồ Việt - Hà Nội - Vietnam

[2]IRISA – Campus de Beaulieu – 35042 Rennes cedex – France

Abstract

The corpus-based constitution of semantic classes is one of the most frequently studied problems in automatic acquisition of lexical semantic information. Most existing works in that area require either that the corpus they exploit use a specialized language, or that they be very large. We describe a generic method for the building of semantic classes from non-specialized, medium-sized corpora, that first splits the studied corpus into topically homogenous ones and then builds separate sets of semantic classes on each of those subcorpora. That approach lets us reduce the importance of polysemy in the studied texts, thus easing the statistical study of word uses. Moreover, a same word can appear with distinct meanings in the various produced sets of semantic classes, which reveals its polysemy potential, and those senses are specified by the knowledge of the topic in which they appear ; that is why we talk about “topicalized” semantic classes. This paper introduces original data analysis techniques that let us achieve fairly good semantic word classifications, although the produced classification trees still require a manual intervention to extract usable semantic classes.

Résumé

La constitution automatique à partir de corpus de classes sémantiques constitue l'une des tâches les plus couramment étudiées en acquisition automatique d'informations lexicales sémantiques. La plupart des travaux existant dans ce domaine exigent soit que les corpus étudiés emploient une langue spécialisée, soit qu'ils soient de taille très importante. Nous proposons une méthodologie générique de construction de classes sémantiques sur des corpus non spécialisés de taille relativement restreinte en découpant ceux-ci en sous-corpus thématiquement homogènes, puis en construisant sur chacun de ces sous-corpus un ensemble de classes sémantiques. Cette procédure permet de réduire la polysémie au sein des textes étudiés, ce qui facilite l'étude statistique des comportements des mots. En outre, un même mot peut révéler des sens différents dans les divers ensembles de classes sémantiques produits, ce qui révèle ses possibilités de polysémie, et ces sens sont spécifiés par le thème au sein duquel ils s'expriment ; c'est pourquoi nous parlons de classes sémantiques « thématisées ». Nous introduisons plusieurs méthodes originales d'analyse statistique de données, qui nous permettent d'aboutir à des classifications sémantiques de mots d'une assez grande pertinence, même si une intervention manuelle reste nécessaire pour extraire des classes sémantiques exploitables des arbres de classification construits.

Mots-clés : classes sémantiques, acquisition sur corpus, thèmes, analyse statistique de données.

1. Introduction

À mesure que progresse la recherche en traitement automatique des langues, les problématiques abordées (résumé automatique, recherche d'information textuelle, etc.) traitent de plus en plus de la sémantique des textes, et requièrent en conséquence des informations sur les sens des mots employés dans ceux-ci, telles qu'on peut en trouver dans des lexiques sémantiques. On considère généralement que les lexiques spécialisés, en fournissant une

information sémantique « ciblée » pour l'étude de textes d'un domaine donné, permettent une exploitation optimale de ces textes. Malheureusement, il n'existe pas de lexique adapté à chaque texte et besoin applicatif, et le coût financier de leur acquisition ou de leur construction manuelle dépasse souvent ce que la tâche entreprise peut justifier. D'où l'intérêt soulevé depuis le début des années 1990 par la possibilité d'apprendre de manière automatique à partir de corpus textuels les sens des mots qu'ils emploient (Hindle, 1990 ; Hearst, 1992). Outre sa rentabilité économique, cette approche permet d'assurer l'adaptation des ressources sémantiques acquises aux textes étudiés, pouvant potentiellement rendre compte non seulement des sens particuliers des mots dans des corpus spécialisés, mais aussi d'usages idiosyncratiques dans des textes non spécialisés. La problématique la plus couramment traitée dans l'acquisition sur corpus d'informations lexicales sémantiques est celle de la construction de classes sémantiques rassemblant des mots de sens proches ; il s'agit également de la question que nous abordons dans cet article. Certains auteurs (Hearst, 1992; Riloff et Shepherd, 1999) proposent de réaliser cette tâche en repérant des motifs linguistiques définis *a priori* qui rapprochent dans le texte des mots de sens proches, tels qu'une énumération par exemple. D'autres (Hindle, 1990 ; Grefenstette, 1993 ; Pereira *et al.*, 1993) comparent les contextes d'apparition des mots pour rassembler ceux employés de manière similaire dans les textes, faisant l'hypothèse que cette proximité d'usage implique une proximité sémantique. Nous définissons pour notre part, dans la lignée des travaux linguistiques de Z. Harris (Harris, 1968), une classe sémantique comme un ensemble de mots interchangeables dans un texte sans affecter la cohésion de celui-ci. Les techniques que nous développons se rapprochent en conséquence naturellement de la seconde catégorie de travaux mentionnée. Elles se distinguent pourtant des recherches existantes par la manière dont elles traitent deux des principaux obstacles à la construction de classes sémantiques par comparaison d'usages de mots : la variabilité lexicale et la polysémie.

Ce que nous nommons variabilité lexicale est la possibilité d'exprimer une même idée d'un grand nombre de manières différentes, en employant des mots différents. Cela constitue naturellement un écueil important pour la comparaison d'usages de mots, car une similarité sémantique entre usages peut être « camouflée » par une différence de forme. Afin de surmonter cette difficulté, de nombreux travaux se restreignent à l'étude de corpus spécialisés (comptes rendus d'opérations médicales, documentations techniques, etc.) (Faure et Nédellec, 1998 ; Bouaud *et al.*, 2000), dans lesquels la langue tend à être plus « rigide », au sens où il existe souvent une manière consacrée d'exprimer les idées-clés et où l'élégance linguistique n'est pas une préoccupation majeure. D'autres (Lin et Pantel, 2001) ne posent pas de contraintes de spécialisation sur la langue du corpus étudié, mais font alors appel à des méthodes statistiques requérant pour fonctionner des quantités de données très conséquentes (50 à 100 millions de mots pour le travail cité).

La polysémie, ou possibilité pour un mot d'être employé avec plusieurs sens distincts, constitue également une source de difficulté importante car elle implique que les informations rassemblées à partir d'un corpus pour définir l'usage d'un mot ne caractérisent pas toutes le même sens, ce qui ajoute beaucoup de « bruit » aux observations. Là encore, une réponse consiste à travailler sur des corpus spécialisés, au sein desquels les sens de mots sont typiquement beaucoup plus stables. Il est également possible de considérer que dans un corpus non spécialisé de taille suffisante, la masse de données disponible permet de « lisser » ce bruit ; les auteurs de (Lin et Pantel, 2001) présentent même une méthode statistique permettant de faire apparaître simultanément un même mot dans plusieurs classes, reflétant ainsi sa polysémie. Il est en revanche impossible de savoir dans ces conditions ce qui dans le

texte engendre l'apparition d'un sens plutôt qu'un autre pour un usage donné d'un mot : on connaît ainsi l'existence des divers sens possibles, mais il n'est pas possible de les exploiter directement.

Nous souhaitons pour notre part proposer une méthodologie générique d'acquisition de classes sémantiques capable de fonctionner sur des corpus « ordinaires », non spécialisés et de taille modérée. Le corpus que nous avons employé pour nos expériences est représentatif de cette exigence : il s'agit d'archives du mensuel *Le Monde diplomatique,* échelonnées sur 14 ans et rassemblant environ 11 millions de mots. La langue employée dans ce corpus est particulièrement riche, présentant une grande variabilité lexicale, et les thèmes abordés sont très variés, donnant lieu à l'expression d'une polysémie importante. Le corpus a été étiqueté morphosyntaxiquement et lemmatisé, mais nous avons choisi de ne pas avoir recours à une analyse syntaxique, par souci de portabilité de la méthode employée.

Le problème de la polysémie est traité en construisant plusieurs ensembles de classes sémantiques « thématisées », correspondant chacun aux sens pris par les mots lors de l'évocation dans le corpus d'un certain thème, en faisant l'hypothèse que ce filtrage thématique permet de stabiliser les sens des mots. Pour cela, le corpus est découpé en sous-corpus thématiques grâce à un système que nous présentons rapidement à la section 2, puis l'acquisition de classes sémantiques est réalisée indépendamment sur chacun de ces sous-corpus. Malgré la réduction de polysémie espérée, les sous-corpus thématiques restent toujours non spécialisés, et présentent une complexité de langue comparable à celle du corpus d'origine. Cette propriété, combinée à leur petite taille (les sous-corpus que nous extrayons rassemblent quelques centaines de milliers de mots), rend particulièrement aigu le problème de variabilité lexicale.

C'est pourquoi nous avons été amenés à mettre au point un système de construction de classes sémantiques en deux étapes. Dans un premier temps, nous calculons en employant le corpus intégral du Monde diplomatique une « distance sémantique » entre mots approximativement révélatrice des ressemblances de sens entre eux (section 3) ; cette mesure est naturellement assez fortement bruitée, notamment par les nombreux cas de polysémie présents dans ce corpus ni spécialisé, ni thématiquement cohérent. Elle constitue néanmoins un outil assez efficace pour enrichir, dans un second temps, l'analyse des contextes d'emploi des mots étudiés sur chacun des sous-corpus thématiques considérés (section 4). En effet, en autorisant la détection de similitudes sémantiques entre voisinages employant des mots distincts mais de sens proches, cette première évaluation de la proximité sémantique entre mots permet une généralisation des observations effectuées sur les données textuelles, et revient donc à « simuler » la disponibilité d'une quantité de données plus importante.

Nous présentons à la section 4.3 les résultats qu'il est possible d'obtenir grâce à la méthode décrite, avant de conclure, section 5, par un aperçu des apports de ce travail et des évolutions qu'il suggère.

2. Découpage du corpus en sous-corpus thématiques

Afin de découper le corpus du Monde diplomatique en sous-corpus thématiques, nous employons le système FAESTOS (Fully Automatic Extraction of Sets of keywords for TOpic characterization and Spotting), présenté dans (Rossignol et Sébillot, 2003). En se fondant exclusivement sur la distribution des mots du corpus sur ses paragraphes, FAESTOS construit de manière totalement automatique un ensemble de classes de mots-clés, chacune représentative d'un des thèmes majeurs abordés dans le corpus. Par exemple, la classe du

thème « nouvelles technologies » pourra rassembler des mots tels que *ordinateur*, *réseau*, *télécommunication*, *etc*.

Par un critère de cooccurrence de mots-clés dans un même paragraphe de texte, ces classes permettent de détecter les occurrences dans le corpus des thèmes qu'elles caractérisent. Nous employons cette capacité afin de regrouper pour chacun des thèmes extraits l'ensemble des paragraphes détectés comme l'abordant, qui constituent un sous-corpus thématique. Les sous-corpus ainsi construits rassemblent de quelques dizaines de milliers à quelques centaines de milliers de mots. Bien que la méthode présentée aux sections suivantes permettent de travailler sur des sous-corpus de relativement petite taille, elle ne permet pas d'obtenir de résultats intéressants en-dessous d'une centaine de milliers de mots ; nous sommes donc limités à une quinzaine de thèmes prédominants, engendrant des sous-corpus de taille supérieure à ce seuil, parmi les quelque 40 mis au jour par FAESTOS.

3. Classification sémantique des mots sur l'ensemble du corpus

Dans un premier temps, notre objectif est de réaliser une classification sémantique des mots sur l'ensemble du corpus, sans aucune préconnaissance sémantique. Étant donné l'usage devant être fait des résultats de cette étape, nous nous intéressons à la classification des principales catégories de mots pouvant servir d'indices lors de la comparaison des contextes d'usages de deux mots à l'étape suivante : noms et noms propres, adjectifs, adverbes, verbes et nombres. Chacun de ces types fait l'objet d'une classification séparée, aboutissant à la définition d'une « distance sémantique » entre mots particulière.

Nous réalisons cette classification par une méthode assez classique en construction sur corpus de classes sémantiques, utilisant une caractérisation ensembliste regroupant de manière non structurée pour chaque mot à classer tous les mots apparaissant à proximité d'une de ses occurrences (section 3.1). La mesure de similarité employée est directement adaptée du classique indice de Jaccard (section 3.2), et nous permet de construire, pour chaque type de mot étudié, un arbre de classification, sur lequel est définie une ultramétrique simple (section 3.3). C'est cette métrique qui est ensuite exploitée comme préconnaissance lors de la construction de classes sémantiques sur des sous-corpus thématiques.

3.1. Représentation des usages des mots

Chaque mot étudié est caractérisé par un ensemble regroupant les mots apparaissant dans au moins n voisinages de ses occurrences. Le paramètre n permet de se limiter aux mots dont il est relativement sûr que leur apparition à proximité du mot-cible n'est pas absolument fortuite ; nous avons empiriquement établi sa valeur à $n = 2$, ce qui permet un « filtrage » élémentaire au prix d'une perte d'information minimale. Les voisinages au sein desquels les mots indices sont collectés sont définis comme des fenêtres de p_d positions à droite et de p_g positions à gauche d'une occurrence de mot. Les valeurs de p_d et p_g sont variables en fonction du type de mots pour lequel nous souhaitons effectuer la classification : pour les adjectifs et adverbes, nous employons les valeurs $p_g = p_d = 1$ (recherche d'une « tête » directement adjacente), pour les nombres, $p_g = 1$ et $p_d = 2$ (cette seconde valeur permettant la présence d'un adjectif ou de la préposition *de* entre un nombre et le nom qu'il caractérise éventuellement), et pour les noms, noms propres et verbes, p g = p d = 3 (les groupes nominaux et verbaux étant typiquement plus « étendus » que les précédents). Ces valeurs ont été établies empiriquement à partir de nombreuses expériences, ainsi que d'après le travail de L. Audibert sur les tailles de voisinage optimales pour la désambiguïsation sémantique (Audibert, 2003) ; nous discutons plus amplement leur choix dans (Rossignol, 2005).

3.2. Mesure de similarité

Cette représentation ensembliste des usages des mots a donné lieu à plusieurs expériences de classification employant diverses mesures de comparaison entre ensembles d'attributs binaires. La plupart de celles-ci ne permettent d'obtenir que des résultats de qualité moyenne, et les plus satisfaisants sont en définitive obtenus grâce à un simple indice de Jaccard normalisé par une méthode que nous avons mise au point dans le but de limiter les effets de « masse de données ». On constate en effet en réalisant une classification ascendante hiérarchique employant un indice de Jaccard brut que certains mots, notamment parmi les plus fréquents, tendent à présenter des valeurs de similarité relativement fortes avec presque tous les autres ; cela induit sur la classification ce que l'on pourrait nommer un « effet trou noir », où tous les objets s'agglutinent autour d'un même noyau de classe très dense au lieu de former des classes clairement distinctes dans l'arbre. Le principe de la méthode mise au point pour éviter cet effet est de normaliser la matrice de similarité afin de ramener dans une même échelle de valeurs toutes les lignes et colonnes de celle-ci.

À cette fin, nous proposons une variante de l'opération statistique de « centrage et réduction » d'une matrice, par laquelle ses valeurs subissent une translation et une mise à l'échelle destinée à rendre leur moyenne nulle et leur écart type unitaire. La variante consiste à centrer et réduire chaque ligne de la matrice — et donc chaque colonne, la matrice de similarité devant rester symétrique. Nous présentons ici rapidement les formules employées pour remplir cet objectif de centrage et réduction ligne par ligne et colonne par colonne d'une matrice, sans détailler les calculs permettant d'aboutir à leur expression. On pourra se référer pour plus de détails à (Rossignol, 2005).

Toutes les modifications réalisées doivent impérativement respecter la propriété intrinsèque de symétrie de la matrice de similarité ; ainsi, si c_i désigne le facteur de translation ajouté aux valeurs de la *i*ème ligne de la matrice afin de la centrer, il devra également être ajouté pour assurer cette symétrie aux valeurs de la *i*ème colonne. Si $M = (m_{ij})$ désigne la matrice originale ($0 \leq i, j \leq n$) et $M' = (m'_{ij})$ la matrice centrée ligne par ligne et colonne par colonne correspondante, M' est donc définie par :

$$\forall (i,j) \in \{1..n\}^2 \quad m'_{ij} = m_{ij} + c_i + c_j \tag{1}$$

Grâce à cette expression et à la propriété souhaitée de centrage de M' , nous aboutissons à l'expression suivante pour les c_i :

$$\forall i \in \{1..n\} \quad c_i = \frac{\bar{m}}{2} - \bar{m}_i \tag{2}$$

où $\bar{m}$ représente la moyenne de tous les m_{ij} , et $\bar{m}_i$ la moyenne des valeurs de la *i*ème ligne. Les (m'_{ij}) étant ainsi définis, nous appelons maintenant $M'' = (m''_{ij})$ la matrice résultant de la réduction ligne par ligne de M', toujours en maintenant sa propriété de symétrie. La réduction est effectuée à l'aide de facteurs multiplicatifs k_i dont chacun doit, comme précédemment, être appliqué à la fois à la *i*ème ligne et à la *i*ème colonne.

M'' est donc définie par :

$$\forall (i,j) \in \{1..n\}^2 \; m''_{ij} = k_i k_j m'_{ij} \tag{3}$$

Il ne nous a pas été possible de donner une solution explicite directe pour le calcul des k_i, mais nous avons montré que la suite $k_{i,j}$ suivante permet leur calcul itératif :

$$\begin{cases} k_{i,0} & = & 1 \\ k_{i,n+1} & = & \frac{1}{2}\left(k_{i,n} + \sqrt{\frac{n}{\sum_{j=1}^{n}\left(k_{j,n}.m'_{ij}\right)^2}}\right) \end{cases} \tag{4}$$

Cette suite converge très rapidement : pour les tailles de données usuelles (de l'ordre du millier d'objets classés), elle permet le calcul de valeurs approchées à 10^{-6} près des k_i en une vingtaine d'itérations. L'application à la matrice *M'* des coefficients de mise à l'échelle k_i présente bien sûr l'inconvénient de lui faire perdre sa propriété de centrage. C'est pourquoi nous procédons en pratique à la répétition des deux procédés de centrage et réduction ligne par ligne jusqu'à stabilisation à ε près des valeurs de la matrice, ce qui se produit en général, suivant la rigueur de l'estimation souhaitée, après quelques dizaines d'itérations.

3.3. Exploitation de la mesure de similarité

La matrice de similarité entre mots centrée et réduite selon le principe exposé à la section précédente est exploitée par un algorithme simple de classification ascendante hiérarchique par lien moyen (*i.e.* la similarité entre deux classes est égale à la moyenne des similarités entre leurs objets). Sur notre corpus de 11 millions de mots, la méthode est appliquée séparément à tous les noms, noms propres, verbes, adjectifs, adverbes et nombres apparaissant plus de 100 fois. À ce seuil, la plupart des regroupements opérés aux premiers niveaux des arbres construits sont pertinents, mais il est courant que des groupes de mots sémantiquement proches à un niveau plus général se trouvent « éclatés » en plusieurs endroits d'un arbre (les classifications ayant la meilleure cohérence globale sont obtenues pour un seuil de 500 occurrences environ). Néanmoins, nous retenons la valeur de 100 occurrences car notre objectif est d'obtenir une première information couvrant le plus grand sous-ensemble possible du vocabulaire du corpus.

La « distance sémantique » entre mots est définie pour chacun des types étudiés comme une ultramétrique sur l'arbre de classification correspondant : la distance *d* entre deux mots est le logarithme à base 2 du cardinal de la plus petite classe de l'arbre qui les rassemble. Cette formule donne la hauteur dans l'arbre à laquelle se trouverait le nœud effectuant la fusion permettant de rassembler les deux mots comparés si l'arbre était binaire et équilibré : il s'agit ainsi d'une manière simple de refléter de manière numérique l'information structurelle représentée par l'arbre. Les valeurs ainsi obtenues sont normalisées à 1 afin d'être exploitables indifféremment comme distances ou proximités (par différence à 1).

4. Classification sémantique des noms sur un sous-corpus thématique

Afin de calculer une similarité d'usage entre mots sur un sous-corpus thématique, il convient de mettre au point une méthode permettant d'exploiter au mieux le peu de données textuelles disponibles. De ce point de vue, la technique employée à la section précédente est clairement sous-optimale, puisqu'en rassemblant dans un ensemble tous les voisins d'un mot, nous perdons toute l'information concernant les collocations de ces voisins entre eux. Afin de conserver cette information, nous avons fait le choix de considérer comme objet d'étude non pas le mot, mais chaque voisinage considéré individuellement. Nous définissons une similarité entre voisinages (section 4.1), à partir de laquelle les similarités entre mots sont ensuite calculées par une technique d'échantillonnage aléatoire que nous avons développée afin de permettre la comparaison non biaisée de mots pouvant présenter des nombres d'occurrences très divers (section 4.2). Cette approche présente l'avantage méthodologique important de distinguer nettement l'opération de comparaison de segments de textes (les voisinages), qui relève d'une approche linguistique, de celle de comparaison globale d'usages de mots (ensembles de contextes), proprement statistique. De plus, en simplifiant le calcul de la similarité, elle permet d'intégrer de manière relativement naturelle la connaissance de la distance *d*, calculée précédemment, dans ce calcul.

L'exploitation des similarités ainsi calculées par une méthode de classification hiérarchique permet la construction d'arbres de classification d'assez bonne qualité (section 4.3), bien qu'une étape de filtrage manuel des résultats soit encore nécessaire pour définir des classes sémantiques à proprement parler.

Nous avons dans la section précédente considéré la classification sémantiques de mots de catégories morphosyntaxiques diverses, guidé en cela par le besoin que nous avions de disposer de premières informations sémantiques pour toutes ces catégories. La construction de classes sémantiques étant dans cette seconde étape une « fin en soi », nous nous limitons par souci de clarté de la présentation à l'étude des seuls noms.

4.1. Représentation des voisinages

La séparation que nous avons réalisée entre comparaison des voisinages et des mots permet toutes les libertés dans la définition d'une mesure de similarité pour comparer les premiers. Celle que nous avons choisie et présentons ici est volontairement simple, afin notamment de maintenir à un minimum les hypothèses réalisées quant aux structures de la langue étudiée. Conformément à notre principe de travail, elle fait usage de la mesure de « distance sémantique » *d* afin d'enrichir la comparaison.

Chaque voisinage est caractérisé par la donnée de l'ensemble des noms, verbes, adjectifs, adverbes, noms propres et nombres apparaissant dans une fenêtre de quatre positions avant et après, respectivement, l'occurrence de mot considérée. La similarité entre deux voisinages est donnée par le simple cardinal de l'intersection de leurs ensembles représentatifs. Le choix de ne pas normaliser cette valeur (par exemple en la divisant par le cardinal de leur union, pour obtenir un indice de Jaccard) s'explique en termes de contrainte sémantique exercée sur un mot par son voisinage : V_1 et V_2 étant les ensembles représentant deux contextes, s'ils ne contiennent qu'un mot chacun et ont ce mot en commun, cela reste un indicateur sémantique faible, même si la coïncidence est de 100 %. En revanche, si V_1 et V_2 rassemblent 5 mots chacun, dont 3 « seulement » en commun, ces mots partagés constituent bien un indice de proximité sémantique potentiellement plus fort.

Afin de prendre en compte la connaissance des rapprochements entre mots acquise sur l'ensemble du corpus à l'étape précédente, nous définissons une mesure de similarité entre deux ensembles de mots V_1 et V_2 consistant à chercher pour chacun des éléments d'un ensemble celui qui lui ressemble le plus dans l'autre ensemble. Mathématiquement, la similarité *s* entre V_1 et V_2 s'exprime donc comme :

$$s(V_1, V_2) = \sum_{a \in V_1} \max_{b \in V_2} (1 - d(a, b)) \quad (5)$$

Cette mesure présente l'inconvénient d'être asymétrique ; la formule réellement employée pour *s* résulte d'une « symétrisation » de celle-ci par moyenne des valeurs réciproques.

4.2. Similarité entre mots à partir des similarités entre leurs contextes

Nous définissons dans cette section le mode de calcul de la similarité *S* entre deux mots m_1 et m_2 apparaissant respectivement n_1 et n_2 fois dans le sous-corpus d'étude à partir d'une similarité *s* définie entre leurs voisinages. Nous notons $v_{ij,\,1 \leq j \leq ni}$ les voisinages du mot m_i .

Une manière « élémentaire » de répondre à cette question consisterait à calculer la moyenne des similarités entre toutes les paires de voisinages de m_1 et m_2 possibles. Néanmoins, cette technique répond mal à la question que nous nous posons, toujours à partir de la définition retenue pour les classes sémantiques : « est-il possible d'utiliser m_2 à la place de m_1 dans les énoncés où ce dernier apparaît, et réciproquement ? ». Pour un voisinage de m_1 donné, la réponse à cette question est positive s'il existe au moins *un* voisinage de m_2 lui ressemblant substantiellement ; il est donc pertinent de prendre en compte pour la comparaison uniquement le voisinage de m_2 ressemblant le plus au voisinage de m_1 considéré. Nous calculons donc, pour évaluer la « remplaçabilité » de m_1 par m_2 , la moyenne des similarités entre les voisinages de m_1 et ceux de m_2 qui leur ressemblent le plus :

$$S(m_1, m_2) = \frac{\sum_{j=0}^{n_1} \max \{s(v_{1j}, v_{2k}) \| 1 \leq k \leq n_2\}}{n_1} \quad (6)$$

La mesure est asymétrique, ce qui requiert une fois de plus une symétrisation par moyenne des valeurs réciproques, mais, plus fondamentalement, elle pose un problème important de sensibilité aux variations de volumes de données entre mots à comparer. La formule choisie limite en particulier très fortement les possibilités de similarité entre m_1 et m_2 si $n_1 \quad n_2$. Afin de contourner cette difficulté, nous n'appliquons la formule donnée par l'équation (6) qu'à des mots présentant exactement le même nombre d'occurrences ; cela est rendu possible par une méthode d'évaluation d'une valeur de similarité entre mots par échantillonnage de leurs représentations.

Cette technique est inspirée des méthodes d'évaluation par échantillonnage aléatoire comme celle de Monte Carlo (Efron et Tibshirani, 1991). Soit deux mots m_1 et m_2 à comparer, caractérisés respectivement par la donnée de n_1 et n_2 voisinages. À partir de ces représentations, nous générons deux populations de *k représentants* de m_1 (resp. m_2), mots

« artificiels » caractérisés chacun par la donnée d'un ensemble de l voisinages tirés au sort (avec remise) parmi les voisinages caractéristiques de m_1 (resp. m_2). L'intérêt de cette opération est de pouvoir créer des représentants comprenant tous le même nombre de voisinages l ; il est ainsi possible de calculer la similarité entre deux représentants de deux mots distincts en employant directement la formule de l'équation (6), sans préoccupation de normalisation. La similarité entre les deux mots d'origine est définie comme la moyenne des similarités entre toutes les paires possibles de représentants de m_1 et m_2 . Les valeurs des mesures de similarité calculées par ce procédé se stabilisent lorsque le produit kl devient nettement supérieur aux nombres d'occurrences respectifs des mots comparés. Lors de nos expériences, travaillant sur des mots apparaissant entre 20 et 1000 fois, nous avons adopté les valeurs $k = l = 70$.

La méthode décrite peut trouver une application dans de nombreux autres contextes d'étude présentant un problème de normalisation, mais ne présente d'intérêt que dans le cas où la similarité entre les objets à comparer dépend d'une similarité définie sur leurs attributs, comme c'est le cas du problème qui nous concerne. Tenter de la mettre à profit pour calculer une similarité entre objets définis, par exemple, par des attributs booléens, ne peut aboutir qu'au même résultat qu'un calcul direct (aux imprécisions introduites par les imperfections d'échantillonnage près).

4.3 Résultats

Nous présentons ici les résultats d'une application des méthodes décrites pour la consti- tution de classes sémantiques sur le sous-corpus du Monde diplomatique correspondant à un thème extrait par Faestos que nous nommons « nouvelles technologies ». Ce sous-corpus rassemble environ 400 000 mots, et nous nous intéressons au quelque 300 noms y apparaissant le plus fréquemment (noms apparaissant plus de 20 fois dans les données considérées, le plus fréquent apparaissant quelque 2 000 fois). Les noms ainsi retenus sont pour beaucoup d'entre eux liés, de manière plus ou moins forte, au domaine des NTIC — qu'il s'agisse de réalités physiques (*dispositif, réseau*), de concepts (*complexité, communication*), d'acteurs (*opérateur, corporation*), *etc*.

La similarité entre noms définie dans cette section permet la construction d'un arbre de classification de relativement bonne qualité, mais dont l'exploitation requiert toujours une intervention manuelle. Il ne nous a en effet pas été possible de définir une technique permettant d'automatiser l'extraction de classes de cet arbre. De plus, il est nécessaire, si nous souhaitons générer des classes sémantiques réellement intéressantes et pouvant constituer une proposition de structuration d'un lexique sémantique, d'étendre l'intervention humaine au-delà de ce simple rôle de sélection, et de permettre un léger filtrage manuel écartant quelques mots parasites empêchant la formation de « bonnes » classes. Sous ces conditions, il est possible de regrouper dans des classes pertinentes environ 60 % des mots étudiés.

Lors de l'étude du corpus correspondant au thème déjà évoqué des « nouvelles technologies », par exemple, quelque 80 classes sont ainsi produites, dont un quart environ requièrent une intervention manuelle telle que décrite ci-dessus. On observe par exemple les classes suivantes :

- {*communication, navigation, tranport, transmission*},
- {*système, programme, dispositif*},
- {*informatique, électronique, biologie*},

- {*groupe, atelier, corporation, entreprise, opérateur, firme, compagnie*},
- {*autoroute, réseau, infrastructure*}.

On peut noter que le filtrage thématique effectué permet bien de faire apparaître des sens de mots spécifiques à un usage particulier correspondant au thème considéré : *autoroute* est à l'évidence compris ici dans le sens d'« autoroute de l'information », et *opérateur* comme un opérateur de télécommunications. Si nous travaillons sur un autre sous-corpus thématique, consacré à la télévision, le groupement obtenu {*programme, programmation, émission*} met à l'évidence en avant un sens de *programme* différent de celui observé précédemment. Un des grands intérêts de la méthodes développée est qu'elle peut permettre la mise au jour de sens particuliers y compris largement minoritaires, en volume, dans le corpus considéré dans sa totalité : ainsi, l'analyse du sous-corpus consacré aux « arts vivants » fait-elle ressortir l'association {*rencontre*, *festival*}, qui correspond à un sens très spécifique de *rencontre* dans un corpus faisant une large place à la narration des rencontres entre chefs d'états, ministres, *etc*.

5. Conclusion

Nous avons présenté dans cet article une méthodologie générale pour l'acquisition automatique, à partir d'un corpus non spécialisé de taille moyenne, de plusieurs ensembles de classes sémantiques correspondant aux divers sens pris par les mots dans les différents thèmes détectés par le système FAESTOS. La méthode proposée est composée de deux étapes :

- la première consiste à réaliser, en employant la totalité du corpus général, un apprentissage de « distances sémantiques » approximatives entre mots, par comparaison des ensembles de mots apparaissant dans les voisinages de leurs occurrences ;
- la seconde fait usage de cette connaissance afin de mesurer les similarités entre les voisinages des occurrences de mots à classer, puis extrapole à partir de ces similarités entre voisinages des similarités entre mots permettant une classification hiérarchique de ceux-ci.

Cette structure nous permet de mettre au jour à partir de corpus de taille assez restreinte des classes sémantiques réellement pertinentes faisant apparaître les différents sens possible d'un mot en spécifiant ces usages par des contextes thématiques : on peut par exemple noter le rapprochement de *réseau* et *autoroute*, révélateur du sens particulier donné à *autoroute* dans le thème « nouvelles technologies » (« autoroutes de l'information »), ou dans le thème des « arts vivants », celui de *festival* et *recontre* (utilisé dans ce sens généralement au pluriel).

Nos recherches ont en outre occasionné le développement de deux méthodes statistiques originales. La première permet la normalisation a posteriori de valeurs de similarité rassemblées dans une matrice, et s'avère un outil précieux pour l'exploitation de formules de calcul de similarité difficilement normalisables *a priori*. La seconde, définie afin de calculer des similarités entre mots à partir des similarités existant entre leurs contextes, constitue selon nous une contribution importante au domaine de l'acquisition automatique sur corpus de classes sémantiques. Elle permet en effet de réaliser une segmentation nette entre le travail proprement linguistique de comparaison de contextes d'apparition de mots, et le processus statistique de généralisation de ces comparaisons à l'échelle de l'ensemble des contextes qui caractérisent l'usage d'un mot. Nous espérons que ce cadre méthodologique occasionnera le développement de méthodes plus avancées pour le calcul de similarités entre contextes, permettant peut-être de dépasser la principale limitation actuelle du système : la nécessité

d'une intervention manuelle sur le résultat de la classification pour la rendre réellement exploitable.

Références

Audibert L. (2003). *Outils d'exploration de corpus et désambiguïsation lexicale au- tomatique*. Thèse de doctorat en informatique, Université de Provence – Aix-Marseille I, Marseille, France.

Bouaud J., Habert B., Nazarenko A., et Zweigenbaum P. (2000). Re- groupements issus de dépendances syntaxiques sur un corpus de spécialité : catégorisation et confrontation à deux conceptualisations du domaine. In Jean Charlet, Manuel Zacklad, Gilles Kassel, et Didier Bourigault, éditeurs, *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, Eyrolles, Paris, France : 275-290.

Efron B. et Tibshirani R. (1991). Statistical Analysis in the Computer Age. *Science*, 253 : 390-395.

Faure D. et Nédellec C. (1998). A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. In *LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications*, Grenade, Espagne.

Grefenstette G. (1993). Automatic Thesaurus Generation from Raw Text Using Knowledge-Poor Techniques. In *Making Sense of Words, 9th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, Oxford, RU.

Harris Z. (1968). *Mathematical Structures of Language*. John Wiley & Sons, New York, NJ, EU.

Hearst M.A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *14th International Conference on Computational Linguistics (COLING 92)*, Nantes, France.

Hindle D. (1990). Noun Classification from Predicate-Argument Structures. In *28st Annual Meeting of the Association for Computational Linguistics (ACL 90)*, Pittsburgh, PA, EU.

Lin D. et Pantel P. (2001). Induction of Semantic Classes from Natural Language Text. In *7th International Conference on Knowledge Discovery and Data Mining (SIGKDD 01)*, San Francisco, CA, EU.

Pereira F.,Tishby N., et Lee L. (1993). Distributional Clustering of English Words. In *31st Annual Meeting of the Association for Computational Linguistics (ACL 93)*, Columbus, OH, EU.

Riloff E. et Shepherd J. (1999). A Corpus-based Bootstrapping Algorithm for Semi-Automated Semantic Lexicon Construction. *Natural Language Engineering*, 5(2) : 147-156.

Rossignol M. et Sébillot P. (2003). Extraction statistique sur corpus de classes de mots-clés thématiques. *TAL (Traitement automatique des langues)*, 44(3) : 217-246.

Rossignol M. (2005). *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*. Thèse de doctorat en informatique, Université de Rennes I, Rennes, France, octobre.