Structuration d'un réseau de cooccurrences lexicales en domaines sémantiques par analyse de textes

Olivier Ferret, Brigitte Grau

Structuration d'un Réseau de Cooccurrences Lexicales en Domaines Sémantiques par Analyse de Textes

Olivier Ferret et Brigitte Grau
LIMSI - CNRS
BP 133, 91403 Orsay cedex, France
[ferret, grau]@limsi.fr

Résumé

Dans cet article, nous présentons une méthode de construction de représentations de thèmes fondée sur la structuration d'un réseau de cooccurrences lexicales. Nous illustrons l'intérêt de l'utilisation d'une segmentation thématique des textes pour réaliser cette structuration, par opposition à un apprentissage réalisé sur le réseau même. Nous tentons aussi de montrer que pour construire la représentation d'un thème, la structuration d'un réseau de collocations donne des résultats plus homogènes que la simple agrégation de segments de texte.

Mots-clés : construction de représentations de thèmes, segmentation thématique, collocations.

1 Introduction

Un certain nombre de travaux relevant du traitement automatique des langues naturelles ont exploité des réseaux de cooccurrences lexicales construits selon les principes présentés dans (Church & Hanks 1990) dans le cadre de la désambiguïsation lexicale (Niwa & Nitta 1994), de l'analyse syntaxique (Hindle & Rooth 1990) ou de la génération (Smadja 1991). De tels réseaux ont l'avantage d'être faciles à construire automatiquement, ce qui permet de traiter des textes sans se limiter à un domaine particulier. En ne codant qu'une proximité entre mots, ils ne constituent cependant qu'une forme faible de connaissances, à la fois peu structurée et peu précise. La cooccurrence entre deux mots peut être aussi bien de nature syntaxique, sémantique ou pragmatique. Il est donc intéressant de pouvoir structurer automatiquement de telles sources de connaissances si l'on veut améliorer les performances des tâches qui les exploitent.

Dans cet article, nous proposons de structurer un réseau de collocations, conçu comme un moyen de rendre compte de la cohésion lexicale, en regroupant ses mots relatifs à un même thème afin de construire une représentation de celui-ci. Ce type de connaissances est impliqué en particulier dans l'identification des thèmes d'un texte.

Opérer ces regroupements en appliquant des techniques d'apprentissage directement sur le réseau ne semble pas constituer une approche pertinente dans la mesure où des collocations fiables ne sont obtenues que pour un petit sous-ensemble de mots très fréquents (Basili et al. 1992). L'approche que nous proposons repose sur l'existence de tels regroupements dans les textes, un texte pouvant être vu comme une succession de segments thématiquement homogènes. Chaque segment ne représente pas à lui seul une description complète d'un domaine car il ne fait que l'évoquer d'une façon particulière et non exhaustive. En revanche, l'analyse de nombreux textes et le recoupement de segments similaires permettent de faire émerger une description générale et stable d'un domaine.

Cette approche rejoint des travaux comme TextNet (Harabagiu & Moldovan 1997) où les textes sont analysés grâce à un noyau initial de connaissances et utilisés comme source d'informations afin d'acquérir d'autres connaissances. Elle se distingue néanmoins des travaux se servant aussi des textes comme source d'informations mais nécessitant d'interagir avec un utilisateur ou de disposer au préalable d'une théorie du domaine bien établie que les nouvelles connaissances se contentent de spécialiser.

2 Description générale

Notre but est de structurer un réseau de cooccurrences lexicales construit à partir d'un vaste ensemble de textes du journal *Le Monde* en créant un niveau permettant de représenter des thèmes, ou domaines sémantiques. Chaque domaine est un nœud faisant référence à des mots du réseau par des liens représentant leur importance dans le domaine. Cette importance est directement liée au nombre d'occurrences du mot dans le domaine et donc, à sa récurrence.

Cette structuration nécessitant de délimiter les segments thématiquement homogènes d'un texte, nous avons développé une analyse thématique exploitant le réseau de collocations. Elle repose sur le calcul d'une valeur de cohésion pour chaque position d'un texte à partir des relations trouvées dans le réseau entre les mots présents dans une fenêtre centrée sur cette position. Les domaines sémantiques sont ensuite construits par agrégations successives de segments relatifs au même thème. Le traitement d'un segment commence par la sélection, à partir de son contenu, des domaines existants les plus proches. Si l'un d'entre eux est suffisamment similaire au segment, les deux sont agrégés de manière à renforcer les mots récurrents, compléter la description du domaine mémorisé et affaiblir l'importance des mots différents. Sinon le segment donne lieu à la création d'un nouveau domaine. De cette façon, des domaines émergent progressivement et se stabilisent après une vingtaine d'agrégations.

Dans une première approche (Ferret & Grau 1998), les mots d'un segment rassemblaient à la fois les mots des textes et les mots du réseau sélectionnés pour évaluer la cohésion des segments (mots inférés). Nous avons cependant constaté la présence d'un bruit important dans les domaines construits, c'est-à-dire de mots non pertinents vis-à-vis du thème du domaine et possédant un poids faible. Fixer un seuil de poids pour filtrer ce bruit ne nous est pas apparu comme une solution satisfaisante sachant qu'une observation fine des domaines nous a aussi montré la présence d'un nombre significatif de mots pertinents de faible poids. Il nous est apparu en revanche que la cohésion thématique des mots inférés était plus grande que celle des mots des textes et leur poids globalement plus élevé.

Aussi, nous avons effectué une seconde expérimentation avec des segments de texte ne contenant que des mots inférés. Les termes explicites n'ont plus alors qu'un rôle de déclencheur. Nous présentons dans cet article les résultats de cette expérimentation, réalisée comme précédemment sur environ 6000 dépêches de l'AFP, et nous les comparons avec ceux de la première expérimentation.

3 Le réseau de collocations

3.1 Le pré-traitement des textes

Pour la tâche considérée, il est important de caractériser les textes par leurs mots significatifs sur le plan de la différenciation thématique. Nous n'avons ainsi retenu que la forme canonique de leurs mots dits pleins, c'est-à-dire les noms, les verbes et les adjectifs. Nous avons éliminé les adverbes, catégorie à l'intérêt très inégal, tous les mots grammaticaux, les noms propres et abréviations ainsi que les verbes auxiliaires ou modaux.

Cette opération est réalisée par une chaîne de traitement ayant pour point de départ des textes sous forme ASCII avec un balisage SGML. Les textes sont d'abord segmentés à l'aide du segmenteur MtSeg du projet Multext. Celui-ci permet en particulier d'identifier les noms composés, lesquels sont particulièrement significatifs puisqu'a priori peu polysémiques. Une liste des 2300 noms composés les plus fréquents sur 11 ans du journal *Le Monde* a été constituée à l'aide d'INTEX (Silberztein 1993) et a été incorporée au segmenteur. Après segmentation, les textes sont étiquetés à l'aide de l'étiqueteur TreeTagger (Stein & Schmid 1995) afin de lever l'ambiguïté sur la catégorie lexicale des formes fléchies et obtenir la forme canonique des mots. Le pré-traitement des textes se termine par la sélection des mots qui représenteront les textes.

3.2 Construction du réseau

Le corpus utilisé pour construire notre réseau de collocations se compose de 24 mois du journal *Le Monde* répartis entre les années 1990 et 1994, ce qui représente un peu plus de 39 millions de

mots. Le calcul des cooccurrences a été réalisé à partir de la méthode décrite dans (Church & Hanks 1990). Le corpus a été pré-traité au préalable selon la procédure présentée au paragraphe précédent, ce qui a conduit à ne retenir qu'environ 37% des mots. L'évaluation des cooccurrences est réalisée en faisant glisser selon un incrément de 1 mot une fenêtre d'une taille de 20 mots sur les textes du corpus. À chaque position de la fenêtre, on enregistre les cooccurrences entre le mot de tête et les autres mots la fenêtre. La fenêtre respecte la délimitation des textes étant donné que deux textes adjacents n'abordent généralement pas les mêmes thèmes. Par ailleurs, nous n'avons pas conservé l'ordre au sein des cooccurrences.

Ces deux derniers points ont été dictés par notre tâche finale. Celle-ci nous a également guidé dans le choix de la taille de la fenêtre. Pour segmenter des textes sur un critère thématique, il est nécessaire de disposer de connaissances à la fois sémantiques et pragmatiques. Ces connaissances nous permettent d'interpréter les relations existantes entre les propositions. Pour capturer ces connaissances au travers d'un réseau de cooccurrences lexicales, il est donc nécessaire que la fenêtre couvre au moins l'équivalent de deux propositions. La limite supérieure est imposée quant à elle à la fois par des limites techniques, le nombre de cooccurrences croît très fortement avec la taille de la fenêtre, et par la granularité des unités que l'on souhaite distinguer. Une taille de 20 mots, après pré-traitement, s'est avérée le meilleur compromis.

mot1	mot2	occurrences	cohésion
imprimante	ordinateur	13	0,227
bateau	voilier	125	0,224
prêtre	curé	44	0,209
chirurgien	hôpital	87	0,195
policier	cambriolage	41	0,190
chômage	emploi	1985	0,167
prendre	racine	120	0,110
collision	franc	7	0,076

Table 1 : Exemples de cooccurrences

À la suite du calcul des cooccurrences, nous opérons une sélection afin de ne retenir que les plus significatives. Seules les cooccurrences de fréquence supérieure à 5 sont conservées, ce qui représente à peu près 1/3 de celles présentes initialement. On aboutit ainsi à un réseau formé de 31000 mots liés par quelques 14 millions de relations. Comme dans (Church & Hanks 1990), nous avons adopté une estimation de l'information mutuelle comme mesure de la cohésion entre deux mots. La taille finie du corpus utilisé pour constituer un tel réseau permet de normaliser cette mesure par l'information mutuelle maximale relative au corpus. Celle-ci est donnée par la formule suivante :

$$I_{max} = \log_2 N^2 (T_f - 1), \text{ avec}$$

N, la taille du corpus et T_f, la taille de la fenêtre.

La table 1 donne quelques exemples de co-occurrences couvrant le spectre des valeurs de cohésion. Ces exemples permettent de constater que le réseau rend compte de relations entre les mots à la fois lexico-syntaxiques (prendre-racine), sémantiques (bateau-voilier ou chômage-emploi) et pragmatiques (policier-cambriolage). Ils montrent également la présence de bruit (collision-franc).

4 La segmentation des textes

Dans la lignée de travaux sur la segmentation thématique des textes tels que (Kozima 1993), nous faisons l'hypothèse que la cohésion lexicale d'un texte est représentative de sa cohérence thématique et que de ce fait, les zones de faible cohésion lexicale sont assimilables à des zones de changement de thème. La méthode de segmentation que nous proposons comporte donc deux étapes. Nous évaluons d'abord la cohésion des différentes parties du texte à segmenter. Nous exploitons ensuite les ruptures significatives de cette cohésion afin de détecter les changements thématiques et créer des segments.

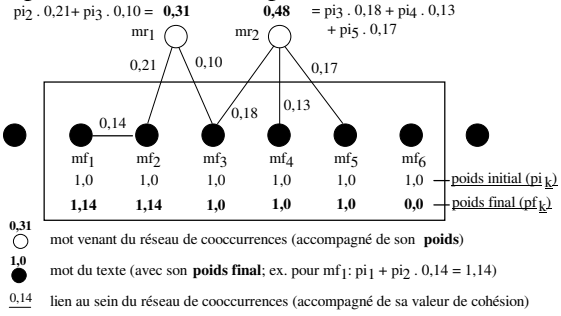

Figure 1 : Calcul du poids des mots

4.1 Calcul de la cohésion d'un texte

L'évaluation de la cohésion d'un texte est réalisée selon des principes proches de ceux mis en œuvre par Kozima mais en utilisant une source de connaissances différente comme référence sur la cohésion lexicale. Cette évaluation consiste à faire glisser une fenêtre d'une taille fixe sur le texte à segmenter et à calculer à chaque station de la fenêtre la valeur de cohésion des mots qui sont présents en son sein à ce moment en utilisant le réseau de cooccurrences présenté ci-dessus. Cette fenêtre est déplacée dans le sens de lecture des textes selon un pas de 1 mot, ceux-ci ayant subi le pré-traitement décrit au 3.1. On obtient ainsi une valeur de cohésion pour chaque position du texte.

Le calcul de la cohésion des mots de la fenêtre se déroule en trois étapes. On sélectionne d'abord les mots du réseau de cooccurrences liés à ceux de la fenêtre et suffisamment proches d'eux sur le plan thématique. Cette proximité est déterminée par le nombre de liens qu'un mot du réseau entretient avec ceux de la fenêtre. En l'occurrence, un mot du réseau est sélectionné s'il est lié à au moins 2 mots de la fenêtre.

Chaque mot retenu, aussi bien venant du réseau que de la fenêtre, se voit ensuite attribuer un poids. Celui-ci est égal au poids initial du mot auquel on ajoute la contribution de l'ensemble des autres mots sélectionnés auxquels il est lié. La contribution d'un mot au poids d'un autre est égale à son poids initial, fixé à 1 pour les mots de la fenêtre et à 0 pour les mots du réseau, modulé par la mesure de cohésion de la relation qui le lie à ce mot (cf. figure 1).

La dernière étape est le calcul de la valeur de cohésion associée à la position courante dans le texte. Cette valeur est égale à la somme des poids des mots retenus, chacun d'entre eux étant modulé par la significativité du mot considéré :

$$coh\acute{e}sion(p) = \sum_i signif(mi) \cdot poids(mi)$$

où $poids(m_i)$ est le poids du mot m_i (appartenant à la fenêtre ou ajouté à partir du réseau) calculé selon les principes décrits ci-dessus.

La significativité d'un mot est définie comme dans (Kozima 1993) par son information par rapport à un corpus de référence, en l'occurrence celui utilisé pour construire le réseau de cooccurrences :

$$signif(m) = \frac{-\log_2(f_m / T_c)}{-\log_2(1/T_c)} \in [0,1]$$

où f_m est le nombre d'occurrences du mot m dans le corpus de référence et T_c est la taille du corpus de référence.

Ce calcul de la cohésion textuelle repose sur l'hypothèse suivante : plus le nombre de mots de la fenêtre relatifs au même thème est grand et plus le nombre de liens que ceux-ci entretiennent dans le réseau de cooccurrences, soit directement, soit par l'intermédiaire des mots ajoutés, est lui aussi grand, ce qui se traduit par une valeur importante de la valeur de cohésion.

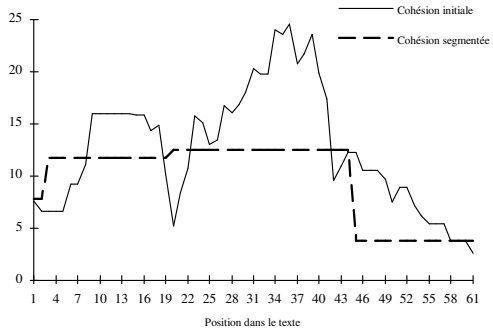

Figure 2 : Courbes de cohésion initiale et segmentée obtenues pour un texte

La figure 2 montre le résultat du calcul de la cohésion pour la totalité d'un texte (cohésion initiale). On y voit apparaître en particulier trois zones (1-20, 21-42, 42-61) correspondant à l'évocation de trois scènes distinctes.

4.2 Délimitation des segments

Pour construire des segments à partir de la courbe de cohésion, nous opérons d'abord un lissage de la courbe afin de faciliter la détection des extrema. Cette opération est réalisée en déplaçant une fenêtre sur le texte et en remplaçant la valeur de cohésion associée au centre de la fenêtre par la moyenne des valeurs de cohésion de toutes les positions incluses dans la fenêtre. Après ce lissage, le calcul de la dérivée de la courbe permet de localiser les maxima et les minima. Compte tenu de l'interprétation associée à la cohésion, les minima sont supposés correspondre à des changements thématiques. Un segment est donc caractérisé par la séquence *minimum – maximum – minimum*. L'étape finale consiste à transformer la

courbe de façon à ce que chaque segment soit représenté par un plateau dont le niveau est déterminé par la valeur du maximum situé entre les deux minima encadrant le segment (cohésion segmentée, figure 2).

5 Les domaines sémantiques

5.1 Représentation des Unités Thématiques et des domaines

Le processus de segmentation permet de délimiter des zones de texte thématiquement homogènes servant à construire des Unités Thématiques (UTs). De façon générale, une UT est la représentation d'un thème élaborée à partir d'un texte. La construction des domaines sémantiques s'effectue par la mémorisation de ces UTs et leur agrégation au fur et à mesure de leur création.

Dans le cadre de cet article, une UT est formée des mots du réseau de cooccurrences les plus fréquemment sélectionnés pour le calcul de la cohésion lors de la délimitation d'un segment. Étant donné que pour chaque position de la fenêtre, on choisit les mots du réseau les plus en rapport avec ceux de la fenêtre, en retenant ceux présents dans une part importante des positions d'un segment, on espère ainsi sélectionner les mots du réseau les plus en rapport avec le segment et donc, avec le thème auquel il fait référence. Plus précisément, ces mots, qualifiés de mots inférés par opposition aux mots figurant explicitement dans les textes, sont les mots du réseau intervenant dans le calcul d'au moins 75 % des valeurs de cohésion se situant à l'intérieur du segment considéré. Au sein d'une UT, les mots sont pondérés par leur significativité.

Un domaine sémantique étant le résultat de l'agrégation de plusieurs UTs, sa structure est identique à celle d'une UT. Il s'agit donc aussi d'un ensemble de mots pondérés. Seule la pondération des mots est différente :

$$poids(m_i) = \frac{nbOcc(m_i)}{nbAgr} \cdot signif(m_i) \cdot \frac{nbAgr^4}{(nbAgr+1)^4}$$

où $nbOcc(m_i)$ est le nombre d'occurrences du mot m_i dans le domaine et $nbAgr$ est le nombre d'agrégations dont résulte le domaine. Le premier facteur rend compte de l'importance du mot par rapport au domaine tandis que le dernier est un modulateur évitant que les domaines récemment créés se trouvent trop favorisés vis-à-vis des plus anciens. Compte tenu de la méthode de construction des UTs, les mots d'un domaine appartiennent nécessairement au réseau de cooccurrences et les domaines constituent donc une structuration de ce réseau selon une dimension spécifique.

5.2 Sélection des domaines

Lorsqu'une nouvelle UT a été construite, il est nécessaire de rechercher en mémoire les domaines qui sont susceptibles de s'agréger avec elle. Pour ce faire, nous réalisons l'équivalent d'un pas de propagation d'activation et nous sélectionnons les domaines les plus activés. L'activation d'un domaine dom_i est donnée par la fonction suivante :

$$activ(dom_i) = \sum_j poids(w_j, dom_i) \cdot poids(w_j, UT)$$

où le premier facteur correspond au poids du mot w_j par rapport au domaine considéré et le second facteur représente le poids du même mot mais vis-à-vis de l'UT à mémoriser.

La sélection des domaines les plus activés s'effectue quant à elle par comparaison à un seuil fonction de la distribution de ces activations : on retient les domaines possédant une activation supérieure à la somme de la moyenne de ces activations et de leur écart-type.

5.3 Similarité et agrégation

Le processus de sélection décrit au paragraphe précédent peut être vu comme une première mesure de similarité, peu élaborée mais applicable largement du fait de son coût raisonnablement faible. Après que le champ des possibilités a été restreint par cette première sélection, il devient possible d'appliquer une mesure de similarité plus complexe afin de déterminer si la nouvelle UT s'agrège à l'un des domaines sélectionnés ou, lorsque la similarité est en dessous d'un seuil fixé a priori, si elle constitue le point de départ d'un nouveau domaine.

Cette mesure de similarité se fonde uniquement sur les mots communs entre un domaine et une UT dans la mesure où la méthode d'apprentissage est source d'un bruit important, même si notre objet est précisément de le réduire. En effet, les

relations présentes dans le réseau de cooccurrences ne sont pas seulement thématiques. Le type des cooccurrences restant implicite, des mots sont donc retenus sur la base d'autres critères que la proximité thématique et représentent une source de bruit du point de vue de notre tâche.

La mesure de similarité entre une UT et un domaine combine l'importance pour chacune de ces deux entités que revêtent leurs mots communs par rapport à l'ensemble des mots qui les forment. Cette importance est elle-même une combinaison du poids de ces mots et de leur nombre d'occurrences. On évite de cette manière d'avoir une forte similarité entre une UT et un domaine ne partageant qu'un petit nombre de mots communs de fort poids de part et d'autre. Chacune de ces deux combinaisons s'effectue en ayant recours à une moyenne géométrique. Plus formellement, on obtient :

$$ratio_{\{d,u\}} = \sqrt{\frac{\sum_c poids(w_c,\{d,u\})}{\sum_t poids(w_t,\{d,u\})} \cdot \frac{\sum_c nbOcc(w_c,\{d,u\})}{\sum_t nbOcc(w_t,\{d,u\})}}$$

$$sim(dom,UT) = \sqrt{ratio_d \cdot ratio_u}$$

où l'indice c fait référence aux mots communs à l'UT (u) et au domaine (d) tandis que l'indice t désigne l'ensemble des mots constituant respectivement l'UT et le domaine. Les mots possédant un poids trop faible dans les domaines (poids inférieur à 0,1) ne sont pas retenus pour le calcul de similarité car ils sont globalement assimilés à du bruit. Le seuil en dessous duquel un nouveau domaine est créé est fixé à 0,25.

L'opération d'agrégation d'une UT et d'un domaine est quant à elle très simple puisque ces entités ont la même structure. Elle consiste pour l'essentiel en la fusion de deux listes de mots pondérés. Le poids d'un mot dans un domaine étant calculé dynamiquement à partir de son nombre d'occurrences, l'agrégation peut être assimilée à une opération additive : si un mot de l'UT n'est pas présent dans le domaine, il y est ajouté avec une occurrence de 1 ; s'il y figure déjà, son nombre d'occurrences est augmenté d'une unité.

6 Expérimentations

6.1 Expérience avec un gros corpus

Nous avons validé notre méthode en l'appliquant sur un mois de dépêches AFP (mai 1994), soit 5949 textes d'une longueur moyenne de 190 mots. À partir d'un ensemble de 7823 UTs issues de la segmentation, 783 domaines ont été construits. 570 résultaient d'au moins 2 agrégations et le plus agrégé rassemblait 351 UTs.

Une étude qualitative des domaines construits montre, sans que cela ait été confirmé par une évaluation rigoureuse (recoupement par plusieurs sujets), que leurs mots les plus représentatifs (poids ≥ 0,1) sont globalement cohérents sur le plan thématique. D'autre part, le contenu de ces domaines se stabilise assez rapidement puisqu'une vingtaine d'agrégations suffisent généralement à faire apparaître des tendances stables. Ceci est illustré par la figure 3 qui rend compte de l'évolution de la tête (les 30 premiers mots) d'un domaine relatif à la justice. Cette évolution est caractérisée par une mesure de similarité tenant compte à la fois du poids des mots et de leur rang au sein du domaine.

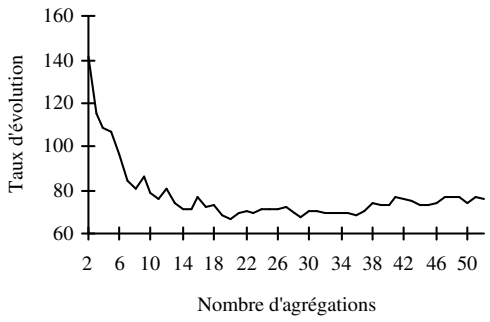

Figure 3 : Évolution de la tête d'un domaine

6.2 Intérêt de la non prise en compte des mots des textes

Dans (Ferret & Grau 1998), nous avons rapporté une expérience (expérience 1) similaire à celle présentée ici (expérience 2) à la différence près que les UTs et les domaines comportaient à la fois les mots venant des segments de texte et les mots inférés à partir du réseau de collocations. On observe qualitativement que les domaines formés

sans les mots des textes sont thématiquement plus cohérents et comportent surtout beaucoup moins de bruit (mots de faible poids). Intuitivement, ce phénomène s'explique par le fait que tous les mots apparaissant dans un segment de texte ne sont pas nécessairement spécifiques du thème abordé par ce segment.

Un certain nombre d'indices plus quantitatifs viennent confirmer ce premier jugement qualitatif. Tout d'abord, on peut noter qu'avec les mots des textes, on forme 3240 domaines à partir de 8601 UTs, ce qui représente un taux d'agrégation nettement inférieur aux 783 domaines mentionnés ci-dessus. La différence concernant le nombre des UTs s'explique par le fait que certains segments ne sélectionnent pas de mots du réseau et ne forment donc pas d'UTs lorsque l'on ne retient que ceux-ci. Cette différence globale se retrouve au niveau des domaines puisque 48 domaines possèdent un nombre d'agrégations au moins égal à 20 dans l'expérience 1 alors que ce chiffre monte à 69 pour l'expérience 2.

Le fait que les mots du texte apportent surtout du bruit par rapport aux mots inférés peut être illustré déjà au niveau des résultats de l'expérience 1. Pour les domaines dont le nombre d'agrégations est au moins égal à 20, domaines supposés stabilisés, on observe ainsi que la proportion de mots significatifs (poids $\geq$ 0,1) est égale à 16,9% pour les mots inférés et à 2,66% seulement pour les mots des textes. On peut voir ce phénomène plus globalement sur la figure 4 qui montre la répartition en fréquence des nombres d'occurrences (assimilables aux poids) des mots des domaines. On peut y constater que les mots inférés fortement récurrents, donc significatifs, sont plus nombreux que leurs homologues venant des textes.

Enfin, la table 2 nous apporte une illustration plus directe de la différence entre les mots des textes et les mots du réseau. On a fait apparaître ici les mots de plus fort poids d'un domaine relatif à la justice après avoir supprimé les mots apparus tantôt comme mots du texte et tantôt comme mots inférés (ces mots sont généralement cohérents entre eux sur le plan thématique). On voit sans conteste que les mots inférés résiduels restent en liaison avec le thème du domaine lorsqu'ils ont un poids suffisant alors que cette liaison est beaucoup plus erratique en ce qui concerne les mots des textes.

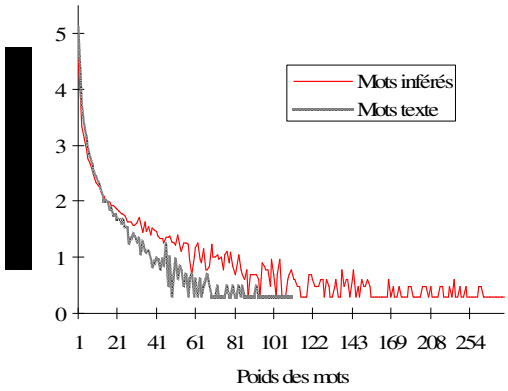

Figure 4 : Fréquence des poids des mots

7 Liaisons avec d'autres travaux

Comme précisé en introduction, la philosophie globale de notre travail se rapproche de celle de TextNet. Il s'agit dans les deux cas de figer des contextes prototypiques à partir de ceux détectés dans les textes. Pour ce faire, on se repose sur un noyau de connaissances permettant de réaliser une analyse minimale des textes.

Mots des textes	Mots inférés
an (0,99)	avocat_général (0,88)
dollar (0,33)	correctionnel (0,88)
avril (0,32)	réclusion_criminel (0,84)
ancien (0,31)	juré (0,84)
dernier (0,30)	cour_d'assises (0,84)
chef (0,27)	condamner_à_mort (0,75)

Table 2 : Mots d'un domaine issus des textes et différents de ses mots inférés et réciproquement

Les contextes sont construits progressivement par le recoupement de passages de texte similaires. La principale différence entre les deux travaux concerne la source initiale de connaissances utilisée. TextNet s'appuie sur WordNet, issu d'une vaste modélisation manuelle, tandis que nous utilisons un réseau de collocations, lequel résulte d'un apprentissage automatique.

En ce qui concerne l'élaboration des domaines sémantiques, notre travail est comparable à celui de Lin (Lin 1997). Les signatures de thème qu'il construit s'apparentent en effet à nos domaines : ce sont des ensembles de mots pondérés en fonction de leur importance et liés au même thème. Plusieurs différences existent cependant entre leur

méthode de création et la nôtre. D'abord, les signatures sont élaborées à partir de textes entiers, en l'absence de tout découpage thématique. Les textes abordant plusieurs sujets sont donc nécessairement source de bruit. Ensuite, la construction des signatures se fait de façon supervisée, les textes étant déjà classés thématiquement. Or un tel classement est rarement disponible. Enfin, la méthode mise en œuvre par Lin n'est pas incrémentale : elle suppose que l'ensemble des textes intervenant dans la construction des signatures soient disponibles en même temps. En revanche, sans aller jusqu'à une hiérarchisation des représentations de thème, elle propose une manière de traiter le problème de la discrimination des domaines très proches.

8 Conclusion

Nous avons développé un système capable de structurer sur le plan thématique un réseau de cooccurrences lexicales. Pour ce faire, nous utilisons la cohérence thématique des textes afin de sélectionner les mots du réseau lié à un thème donné. La construction de la représentation d'un thème s'effectue par le cumul des mots ainsi sélectionnés pour un ensemble de textes abordant ce thème. Nous avons montré que l'on peut construire de cette façon des domaines sémantiques stables et significativement homogènes sur le plan thématique. Nous avons également montré que cette stratégie est préférable à celle consistant à agréger directement les segments constitués des mots des textes. Une observation globale des domaines construits montre cependant que ceux-ci ne sont pas toujours disjoints et qu'il conviendrait donc de les organiser de façon hiérarchique, soit en cours de construction, soit a posteriori. Les algorithmes étudiés devront à la fois être capables d'abstraire les points communs de deux domaines mais également de les diviser dans le cas fréquent où ils regroupent un thème principal ainsi qu'un ou plusieurs thèmes secondaires.

Références

(Basili et al. 1992) R. Basili, M.T. Pazienza and P. Velardi, *Computation lexicons: the neat examples and the odd exemplars*, Proceedings of the 3rd ANLP conference, 1992.

(Church & Hanks 1990) K.W. Church and P. Hanks, *Word Association Norms, Mutual Information, And Lexicography*, Computational Linguistics, Vol. 16-1, 1990.

(Ferret & Grau 1998) O. Ferret and B. Grau, *A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts*, Proceedings of ECAI'98, Brighton, 1998.

(Harabagiu & Moldovan 1997) S.M. Harabagiu and D.I. Moldovan, *TextNet – A text-based intelligent system*, Natural Language Engineering, Vol. 3-2/3, 1997.

(Hindle & Rooth 1990) D. Hindle and M. Rooth, *Structural ambiguity and lexical relations*, Proceedings of DARPA Speech and Natural Language Workshop, Hidden Valley, PA, 1990.

(Kozima 1993) H. Kozima, *Text Segmentation Based on Similarity between Words*, Proceedings of the 31th Annual Meeting of the ACL (Student Session), Colombus, Ohio, USA, 1993.

(Lin 1997) C.-Y. Lin, *Robust Automated Topic Identification*, Doctoral Dissertation, University of Southern California, 1997.

(Niwa & Nitta 1994) Y. Niwa and Y. Nitta, *Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries*, Proceedings of the 15th COLING conference, Kyoto, Japan, 1994.

(Silberztein 1993) M. Silberztein, *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Masson, 1993.

(Smadja 1991) F. Smadja, *Retrieving collocational knowledge from textual corpora. An application: Language generation*, Doctoral Dissertation, Columbia University, 1991.

(Stein & Schmid 1995) A. Stein et H. Schmid, *Étiquetage Morphologique de Textes Français avec un Arbre de Décisions*, Traitement Automatique des Langues, Vol. 36-1-2, 1995.