Extraction non supervisée sur corpus de classes de mots-clefs

Unsupervised extraction of keyword classes from corpus

Mathias Rossignol Pascale Sébillot

IRISA

Campus de Beaulieu, 35042 RENNES Cédex
(mrossign|sebillot)@irisa.fr

Résumé

Nous présentons une méthode non supervisée d'acquisition sur corpus d'ensembles de mots-clefs, c'est-à-dire de mots dont la présence dans un texte est révélatrice de l'apparition d'un thème donné. Les classes ainsi construites sont associées chacune à l'un des principaux thèmes du corpus, et peuvent être employées afin de détecter l'apparition de ce thème dans un paragraphe de texte, par un simple principe de cooccurrence de mots-clefs. Leur génération est réalisée de manière totalement automatique grâce à la combinaison de plusieurs techniques d'analyse statistique de données, et notamment la méthode de classification hiérarchique CHAVL.

Mots Clef

Détection de thèmes, caractérisation de thèmes, classification non supervisée.

Abstract

We present an unsupervised method for the generation from a textual corpus of sets of keywords, that is, words whose appearance in a text is strongly connected with the presence of a given theme. Each of these classes is associated with one of the main themes of the corpus, and can be used to detect its presence in any paragraph of the corpus, by a simple keyword cooccurrence criterion. The extraction of these classes from the textual data is carried on totally automatically thanks to the combination of several statistical data analysis techniques, including the CHAVL *hierarchical clustering method.*

Keywords

Topic detection, topic characterization, unsupervised classification.

1 Introduction

La tâche consistant à caractériser et détecter les occurrences des thèmes abordés dans un texte constitue une problématique récurrente dans le domaine du TAL (Traitement Automatique des Langues) : on la rencontre par exemple en recherche documentaire (afin de trouver un texte " parlant de . . . ") ou dans le cadre d'applications réalisant une analyse sémantique plus fine, afin de désambiguïser le sens de mots polysémiques. Elle a donné lieu à de nombreux travaux, dont une sélection est présentée ici en section 2, qui la couplent pour la plupart avec celle de segmentation du corpus en segments textuels monothématiques.

Il peut néanmoins être préférable de respecter la structuration du texte réalisée par son auteur, notamment dans les cas où l'on souhaite que les segments de textes extraits par la détection soient sémantiquement " autosuffisants " : pour une application de recherche documentaire, par exemple, il n'est pas pertinent de fournir à l'utilisateur une phrase isolée, voire une portion de phrase. Nous avons donc choisi de considérer qu'un texte se trouve naturellement découpé en unités thématiques relativement cohérentes, ses paragraphes, et c'est à cette échelle que le système présenté réalise la détection de thèmes. À partir d'un corpus de texte, nous construisons un ensemble de *classes de mots-clefs*, c'est-à-dire d'ensembles de mots dont la présence dans un texte est particulièrement révélatrice d'un thème donné – par exemple, la classe de mots-clefs associée au thème *agriculture* peut rassembler les mots *agriculteur, blé, champ, ensemençage, ferme, maïs, etc*. Ces classes représentent les principaux thèmes abordés dans le corpus étudié, et permettent la détection de leurs apparitions par un simple critère de cooccurrence : nous considérons que la présence conjointe dans un paragraphe de deux mots-clefs d'une même classe est suffisante pour affirmer que le thème sous-jacent à celle-ci s'y trouve abordé.

La méthode la plus efficace pour réaliser la construction de ces classes consisterait naturellement à faire usage d'un corpus d'apprentissage dont les paragraphes ont déjà été regroupés manuellement par thèmes, et à rassembler les mots les plus spécifiques à chaque collection de paragraphes – c'est-à-dire les mots présentant sur celles-ci des fréquences anormalement élevées eu égard à leur répartition moyenne dans le corpus. C'est par exemple l'approche retenue et dé-

veloppée de manière plus fine dans [4]. Cette méthode supervisée présente néanmoins deux inconvénients majeurs : d'une part, elle nécessite un important travail de préparation des données pour permettre l'apprentissage, dont le coût ne se justifie pas toujours au regard de l'enjeu applicatif, et elle mène d'autre part à faire une présupposition sur les thèmes abordés dans le corpus. Nous souhaitons pour notre part nous affranchir de la nécessité d'une intervention humaine, et faire émerger les thèmes du texte lui-même, ce qui nous assure une adaptation optimale aux données traitées et peut en outre faire apparaître des informations nouvelles, potentiellement contre-intuitives, le concernant.

Le système présenté réalise donc la construction des classes de mots-clefs sans faire usage d'aucune connaissance préalable concernant le corpus, les sens de ses mots, ou les thèmes y apparaissant, et sans requérir d'intervention humaine – nous acceptons en contrepartie de cela la possibilité que la détection de thèmes réalisée par ces classes soit parfois imparfaite ou incomplète.

Considérant que les mots évoquant un même thème ont de fortes chances d'apparaître conjointement dans les mêmes paragraphes, nous faisons usage d'une méthode de classification statistique " classique " rapprochant les mots de répartitions similaires sur le corpus, dont les résultats sont exploités grâce à un algorithme destiné à mieux prendre en compte les spécificités des données traitées. Ce traitement, détaillé en section 4.1, permet l'organisation en classes d'une sélection de mots du corpus choisis selon leur type grammatical (noms et adjectifs) et leur fréquence [1].

Ce mode de sélection est naturellement peu satisfaisant : les noms et adjectifs fréquents du corpus n'ont pas tous vocation à jouer le rôle de mots-clefs, et les classes obtenues sont donc " parasitées " par de nombreux mots non pertinents pour la détection de thèmes. Afin de filtrer ceux-ci, nous ne traitons pas l'intégralité des données en une seule fois, mais réalisons plusieurs classifications sur des extraits aléatoires distincts du corpus, puis ne retenons les associations entre mots que si elles sont observées dans plusieurs de ces classifications partielles (section 4.2). Cela permet d'écarter de la classification finale les mots " thématiquement neutres ", qui ne participent à aucun regroupement stable, et également de résoudre les problèmes de " bruit " subsistant à l'issue du premier traitement. On obtient ainsi un ensemble de classes contenant chacune quelques mots très fortement associés à un thème particulier.

Ce premier résultat est toujours tributaire du mode de sélection initial des mots à classer : nous ne disposons en effet que des mots-clefs les plus fréquents dans le corpus, alors que de nombreux mots plus rares peuvent être fortement porteurs d'information thématique. Nous employons donc ces premières classes afin d'effectuer une première détection de thèmes sur l'ensemble du corpus, et de réaliser à partir de ce résultat un apprentissage de type supervisé prenant en compte les mots trop peu fréquents pour avoir pu être considérés auparavant. Cette dernière étape de construction, décrite à la section 5, permet d'obtenir des classes d'une trentaine de mots en moyenne rendant assez fidèlement compte de l'étendue du " champ lexical " associé à chacun des principaux thèmes du corpus étudié.

Nous présentons à la section 6 les résultats obtenus et les évaluons de deux points de vue : la cohérence intuitive des classes produites et leur efficacité pratique pour la détection de thèmes. Enfin, quelques directions possibles pour une évolution du système sont évoquées en conclusion.

2 Autres approches de la détection de thèmes

Plusieurs travaux se sont intéressés à la détection automatique de thèmes en s'appuyant soit sur des indices linguistiques [10], soit sur des notions telles que la cohésion lexicale [6, 13, 5], la plupart d'entre eux réalisant simultanément la caractérisation de thèmes et une segmentation du texte en fragments thématiquement cohérents. Les objectifs des dernières recherches citées étant plus proches de nos préoccupations, nous présentons ici rapidement leurs principes de manière à mieux situer le travail présenté ici, en ce qui concerne tant les méthodes mises en œuvre que les objectifs poursuivis.

L'un des travaux fondateurs dans le domaine de la détection de thèmes, *TextTiling* [6], réalise le découpage d'un texte en groupes de paragraphes successifs portant sur un même thème en se basant sur une mesure de similarité lexicale entre séquences consécutives de mots. Tous les 20 mots, l'algorithme calcule la ressemblance entre les listes des 100 mots apparaissant à gauche et à droite du point de focus. Un minimum local de cette mesure est considéré comme l'indice d'une zone de changement thématique, et une frontière est alors définie à l'emplacement de la limite de paragraphe la plus proche. Les mots ayant joué un rôle important dans le maintien à une valeur élevée de la mesure de cohérence lexicale entre deux minima sont mis à profit pour caractériser le thème de la région considérée. Si l'algorithme décrit permet en effet une segmentation thématique pertinente à l'échelle de groupes de paragraphes, rien ne garantit que deux zones non contiguës traitant d'un même thème soient caractérisées par des listes de mots identiques, ce qui rend difficile la détection de proximité thématique entre segments non consécutifs.

Cette limitation est surmontée par le couple de traitements SEGCOHLEX / SEGAPSITH, présenté dans [5], qui réalise en outre une segmentation de granularité beaucoup plus fine en calculant la mesure de cohésion lexicale pour chaque occurrence de mot et à l'aide de fenêtres de 10 mots à gauche et à droite du mot cible. Cet indice faisant usage d'une quantité de termes bien plus faible que le précédent, des données supplémentaires sont utilisées pour accroître indirectement la quantité d'information qu'il rassemble : un corpus de 45 millions de mots est employé au préalable afin d'extraire un réseau de collocations, et cette connaissance est exploitée dans la mesure de proximité lexicale

1. L'analyse statistique des répartitions des mots ne peut en effet être significative qu'au-delà d'un certain nombre d'occurrences.

pour réaliser une première segmentation des textes étudiés (module SEGCOHLEX). De cette segmentation, le système extrait des " signatures thématiques " qui forment la base d'une seconde segmentation (module SEGAPSITH) et fournissent une caractérisation indirecte mais – contrairement à *TextTiling* – systématique des thèmes détectés.

Nos objectifs diffèrent de ceux de ces travaux en ce que nous ne nous intéressons pas à une structuration de la lecture séquentielle des textes, mais cherchons à construire une caractérisation permettant de connaître de manière immédiate le thème abordé dans un segment textuel défini *a priori* : en effet, le découpage du texte en paragraphes est intrinsèque aux données étudiées et pré-existe à l'application des traitements que nous définissons. Nous n'effectuons donc pas un parcours analytique linéaire du corpus, mais traitons " en bloc " un ensemble de données statistiques collectées sur celui-ci.

Le travail présenté par G. Salton *et al.* dans [13] aborde une problématique plus proche la nôtre, en ce qu'il travaille également à l'échelle du paragraphe et ne se limite pas à une approche séquentielle du texte : l'objectif en est d'extraire la structure thématique selon laquelle s'organisent les paragraphes d'un article. Les auteurs réalisent dans un premier temps un regroupement de paragraphes successifs selon une méthode similaire à celle employée par TextTiling, puis rapprochent les groupes non contigus selon un critère de similarité lexicale calculé à l'échelle de paragraphes entiers. Ces informations de structure sont utilisées afin de guider la sélection des passages d'un article les plus pertinents en réponse à une requête documentaire. La problématique de caractérisation des thèmes sous-tendant la structure extraite n'est pas abordée, ce qui limite l'intérêt direct de la méthode employée pour le cas qui nous préoccupe. Nous faisons néanmoins usage au cours des traitements développés ici d'une mesure de similarité lexicale entre paragraphes apparentée à celle employée par Salton, quoique reformulée pour l'adapter à notre problématique et aux quantités de données beaucoup plus importantes que nous traitons.

De nombreux travaux, enfin, obtiennent de très bons résultats grâce à l'emploi d'outils plus " raffinés " pour la détection de thème : les auteurs de [3] font ainsi usage de plusieurs modèles statistiques de language (grammaires n-grammes et modélisation par " mémoire cache " [7]) afin de caractériser la forme de texte typique associé à chaque thème. L'apprentissage de ces modèles étant réalisé de manière supervisée sur des ensembles de textes de thèmes connus, il est plus difficile dans ce cas de faire une comparaison avec nos travaux. Nous avons en effet fait le choix de ne pas faire dépendre les résultats obtenus du travail d'un expert, et de faire émerger les thèmes reconnus des données textuelles elles-même.

Avant de procéder à la description du système que nous avons développé, nous présentons en quelques mots le corpus ayant servi de support aux expérimentations menées lors de notre travail.

3 Corpus d'étude

Le corpus employé lors de notre étude est constitué de quinze ans d'archives du mensuel *Le Monde Diplomatique*, dont on n'a retenu que les éléments relativement développés d'un point de vue rédactionnel (articles de réflexion, reportages, analyses de livres ou films . . . , par opposition aux brèves, *etc.*). Cette sélection rassemble 5 704 articles, découpés en quelque 98 000 paragraphes et rassemblant un total de plus de 11 millions de mots.

Ce corpus a été segmenté, étiqueté (étiquetage catégoriel) et lemmatisé grâce aux outils MULTEXT [1] et TATOO [2]. Cet étiquetage nous permet dans la suite de restreindre la liste des mots-clefs potentiels aux types de mots les plus volontiers porteurs d'information thématique, la lemmatisation nous affranchissant pour sa part du bruit statistique introduit par les variations morphologiques.

Ces précisions étant apportées, nous pouvons dans les sections suivantes présenter les différents traitements composant le système développé.

4 Extraction des thèmes

Nous montrons dans cette section comment le système présenté s'organise autour d'une méthode " classique " de classification de données afin de générer un ensemble de classes de mots-clefs à partir des données de répartition des mots sur le corpus d'étude, puis comment un filtrage de ces résultats est effectué afin d'obtenir une collection de classes " minimales " caractérisant les principaux thèmes du corpus, à partir desquelles nous construisons dans la section suivante des classes plus complètes permettant la détection de ces thèmes.

Le principe des traitements présentés ici est issu d'une idée présentée dans [11]. Les auteurs y considèrent que, puisque les mots que l'on cherche à rassembler sont révélateurs d'un même thème, dont leur présence dans un paragraphe est un indicateur fort, il existe une forte probabilité de les voir apparaître conjointement dans les mêmes paragraphes du corpus – ceux qui abordent le thème considéré. L'objectif est donc ici de regrouper les mots présentant une distribution similaire sur les paragraphes du corpus.

Nous utilisons pour cela la méthode de classification hiérarchique CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance du Lien [8])[2], qui construit un arbre de classification des mots à partir de leurs données de répartition. Des classes sont extraites de cet arbre grâce à un algorithme de lecture " intelligente " développé pour mieux prendre en compte les spécificités des données étudiées.

Afin de permettre le filtrage des résultats, ce couple de traitements n'est pas appliqué en une seule fois à l'ensemble des données, mais de manière répétitive sur des extraits aléatoires distincts représentant chacun environ 10 % du

2. Étant donnée notre ignorance du nombre de thèmes, donc de classes, recherché, l'usage d'un algorithme de classification hiérarchique s'impose. Nous avons en revanche pu constater expérimentalement que le choix d'une méthode en particulier n'influait que marginalement sur l'ampleur et la nature des " post-traitements " à effectuer.

corpus total. C'est en confrontant des résultats de ces diverses classifications que le système parvient à ne retenir que les regroupements de mots les plus pertinents.

4.1 Construction et exploitation d'un arbre de classification des mots

La méthode de classification CHAVL est appliquée au tableau de contingence croisant les lemmes des 1 000 noms et adjectifs les plus fréquents du corpus avec 10 000 paragraphes sélectionnés aléatoirement dans le corpus : chaque case du tableau de données employé contient le nombre d'occurrences d'une forme quelconque du lemme considéré dans un paragraphe donné. La mesure employée pour calculer les similarités entre lemmes à partir des lignes de ce tableau est présentée dans [9].

La figure 1 représente une version réduite de l'arbre de classification que l'on obtient à l'issue de ce traitement. Chacun des nœuds de l'arbre correspond à une classe résultant de la fusion des classes correspondant à ses fils, et l'on voit ainsi apparaître à mesure que l'on remonte des feuilles de l'arbre vers sa racine des classes de plus en plus développées et générales.

L'arbre obtenu présente plusieurs caractéristiques rendant son exploitation problématique : d'une part, la classification réalisée n'est pas sans défaut (ainsi le rassemblement de *capitalisme* avec *vie* ne suggère-t-il pas réellement de thème), ce qui s'explique notamment par le fait que le tableau de répartition des mots sur les paragraphes est très creux (99 % de ses cases ont en effet une valeur nulle), et met donc en difficulté la méthode de classification employée. D'autre part, on peut constater que les classes les plus intéressantes pour nous se forment à des hauteurs très diverses dans l'arbre. Il est donc impossible de faire usage de la méthode traditionnelle de lecture de tels arbres, par coupure à un niveau unique (ligne pointillée sur la figure) : si l'on souhaite ici obtenir la classe {*chômage, emploi, chômeur*}, il est nécessaire de sélectionner un niveau proche de la racine (le pointillé) et de perdre des distinctions également intéressantes sur le reste des objets.

Il nous faut donc mettre au point une méthode de lecture capable, d'une part, de choisir les classes les plus pertinentes à différents niveaux de l'arbre, et, d'autre part, d'écarter de la classification les regroupements qui, quoique toujours justifiés statistiquement, ne le sont pas d'un point de vue thématique. Afin d'être capable d'effectuer ce choix, nous définissons une mesure numérique évaluant la qualité d'une classe pour la détection de thèmes. Nous utilisons pour cela une classification réciproque de celle réalisée précédemment, en classant les paragraphes de l'extrait étudié en fonction de la similarité du vocabulaire qu'ils emploient. L'objectif est d'obtenir ainsi un regroupement thématique des paragraphes (en faisant l'hypothèse que deux segments de texte employant les même mots parlent sans doute de la même chose) qu'il sera ensuite possible de confronter à la classification des mots : il s'agit en effet de deux réponses distinctes à la même problématique de détection de thèmes, qui devraient donc coïncider autant que possible.

Nous présentons la technique mise en œuvre pour classer les paragraphes, avant d'exposer la manière dont cette " coïncidence " est quantifiée sous la forme d'une mesure de qualité et exploitée algorithmiquement.

Classification des paragraphes. Pour cette classification, nous avons choisi une mesure de similarité lexicale simple, qui consiste à déterminer le nombre de mots partagés par deux paragraphes pour évaluer leur proximité thématique. Ce principe est affiné en donnant un poids supplémentaire au partage de mots rares, dont l'apparition vraisemblable dans peu de thèmes distincts fait de leur présence dans un paragraphe un indicateur thématique fort. L'importance de chaque mot est donc inversement proportionnelle à son nombre d'occurrences dans le corpus d'apprentissage, et la mesure est normalisée en fonction de la taille des paragraphes comparés. Par conséquent, la similarité entre deux paragraphes A et B est définie par :

$$s\,(A,B) = \frac{1}{\min(n_A,n_B)} \sum_i \frac{\min(a_i,b_i)}{n_i}$$

où $A = (a_i)$ et $B = (b_i)$ sont les vecteurs rassemblant le nombre d'occurrences de chaque mot considéré pour ce calcul dans chacun des paragraphes, n_i est le nombre total d'occurrences du mot i dans les 10 000 paragraphes du corpus extrait, et n_A et n_B les nombres de mots des deux paragraphes.

De par sa simplicité, cette mesure est à même de prendre en compte également les mots les plus rares du corpus et nous permet d'effectuer le calcul pour tous les noms et adjectifs apparaissant au moins 2 fois dans l'extrait étudié, soit 9 000 mots environ, ce grand volume de données exploitées compensant partiellement la relative naïveté de s. La demi-matrice 10 000 x 10 000 contenant les indices de similarité entre paires de paragraphes est traitée par CHAVL pour construire un arbre de classification des paragraphes. L'arbre obtenu, assez bien équilibré, nous permet d'extraire, par coupure à un niveau où la taille moyenne des classes atteint une valeur que nous avons fixée à 12 [3], un ensemble d'environ 600 classes (certaines classes s'éloignant trop de la taille moyenne choisie sont écartées).

Confrontés à l'impossibilité pratique de vérifier manuellement la pertinence d'un arbre de classification rassemblant 10 000 éléments, nous ne disposons pour juger de la qualité de cette partition que d'indicateurs partiels ou indirects : le relativement bon équilibre de l'arbre obtenu est ainsi le signe d'un déroulement satisfaisant du processus de classification, et des validations manuelles effectuées afin de contrôler la cohérence de quelques classes choisies au hasard montrent que celles-ci présentent un degré acceptable, quoique loin d'être parfait, de cohérence thématique.

3. Cette valeur empirique est un bon compromis entre généralisation et consistance thématique des classes.

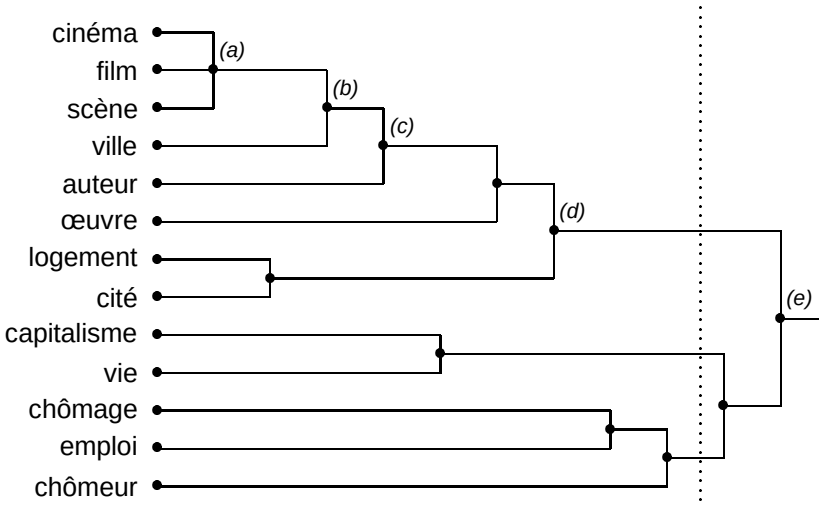

FIG. 1 – *Exemple réduit d'arbre de classification des noms produit par* CHAVL. *La ligne pointillée indique un niveau de lecture traditionnel de tels arbres.*

La classification des paragraphes obtenue constitue un point de référence permettant de diriger l'extraction des classes de mots-clefs de l'arbre de classification des mots.

Extraction de classes de mots-clefs de l'arbre. Les classes de mots-clefs que nous nous proposons de construire ont pour objectif de permettre la détection de l'apparition d'un thème dans un paragraphe selon le principe suivant : si deux mots d'une classe sont présents dans un paragraphe, alors il aborde le thème sous-jacent à celle-ci. On dira dans ce cas que la classe *reconnaît* le paragraphe. Dans le cas idéal d'une classification parfaite, où chacune des classes de paragraphes obtenues précédemment rassemble en effet des paragraphes abordant le même thème, et où chaque classe de mots reconnaît tous les paragraphes abordant " son " thème, et seulement ceux-ci, on se trouve dans une situation où soit une classe de mots reconnaît tous les paragraphes d'une classe, soit elle n'en reconnaît aucun.

Nous définissons une mesure q de qualité d'une classe de mots-clefs potentielle mesurant à quel point la détection qu'elle réalise sur les paragraphes est proche de cette situation. Son mode de calcul est présenté en détail dans [12], et nous n'en exposons ici que le principe général : considérant une classe de mots-clefs $\mathcal{C}$, on calcule pour chacune des classes de paragraphes la proportion de ses éléments qui est reconnue par $\mathcal{C}$. Dans le cas idéal décrit, ces proportions valent toutes soit 0, soit 1. La mesure q est formulée de manière à être d'autant plus élevée que leurs valeurs se répartissent en effet en deux catégories clairement distinctes de valeurs fortes et faibles.

L'intérêt de cette mesure pour la lecture de l'arbre de classification est double : d'une part, elle sert à pointer les classes pertinentes dans l'arbre quel que soit leur niveau ; d'autre part, elle permet d'ignorer certains regroupements de mots effectuées par CHAVL, voire de les modifier. Nous présentons le déroulement de l'algorithme développé à cette fin en nous appuyant sur l'exemple réduit de la figure 1, similaire à l'arbre de classification de 1 000 mots réellement traité. L'algorithme part des feuilles de l'arbre de classification et remonte vers la racine en vérifiant à chaque nœud rencontré si la fusion proposée accroît la valeur de q.

- Si c'est le cas, l'algorithme effectue la fusion et continue l'exploration ascendante de l'arbre avec cette nouvelle classe.
- Sinon, l'algorithme continue à remonter vers la racine mais sans effectuer la fusion. On se retrouve donc avec un ensemble de classes au lieu d'une simple classe de mots-clefs. En (b), par exemple, le programme décide de ne pas effectuer la fusion et poursuit l'exploration ascendante de l'arbre avec un ensemble de 2 classes : {{*cinéma, film, scène*}, {*ville*}}.

Aux nœuds supérieurs, toutes les possibilités de fusion entre membres des ensembles de classes sont testées pour détecter les plus intéressantes en termes d'évolution de q. Par exemple, en (d): {{*cinéma, film, scène, auteur, œuvre*}, {*ville*}} " + " {*logement, cité*} → {{*cinéma, film, scène,auteur, œuvre*}, {*ville, logement, cité*}}
Nous obtenons enfin en (e), racine de l'arbre, la partition : {{*cinéma, film, scène, auteur, œuvre*}, {*ville, logement, cité*}, {*capitalisme,vie*}, {*chômeur, emploi, chomage*}}.
Les classes extraites sont donc différentes de celles produites par CHAVL et sont obtenues à l'issue du parcours sans qu'il soit nécessaire de les chercher à divers niveaux de l'arbre. Une heuristique consistant à détecter les classes " stabilisées " (c'est-à-dire peu susceptibles d'améliorer leur qualité par fusion avec une autre classe) et à les écarter du processus de recherche des meilleures fusions permet de réduire le temps de calcul[4].

4. L'ensemble du système de construction / lecture de l'arbre de classification s'exécute en 20 à 25 minutes sur une machine de type Sun Blade.

L'ensemble de classes ainsi obtenu, quoique d'assez bonne qualité, est encore trop imparfait pour effectuer une bonne caractérisation des thèmes du corpus : comme nous l'avons déjà mentionné, cela est notamment dû au fait que les mots choisis pour la classification le sont uniquement en fonction de leur fréquence. Des mots n'ayant pas vocation à être des mots-clefs viennent ainsi " parasiter " les classes produites, voire occasionner des rapprochements entre mots sans rapport thématique.
Afin de filtrer ce " bruit ", nous confrontons les résultats de plusieurs exécutions de ce premier système sur des extraits distincts, pour ne retenir que les regroupements " stables ".

4.2 Filtrage par confrontation de plusieurs classifications

Notre objectif est de regrouper les mots les plus fréquemment rassemblés dans les différentes partitions des mots obtenues au cours de n exécutions du processus de construction et lecture de l'arbre de classification [5]. Pour ce faire, l'ensemble des partitions est synthétisé sous la forme d'un graphe valué dont les sommets correspondent aux mots classés et sont reliés entre eux par des arcs dont le poids est égal au nombre de fois où les mots ont été regroupés dans une même classe. Une partie du graphe ainsi obtenu est présentée en figure 2, où l'on voit clairement apparaître les classes que nous nous proposons d'extraire.
Cette extraction s'effectue de la manière suivante : sélectionnant dans le graphe l'arc de poids le plus fort, on définit un seuil sur les poids des arcs du graphe en-deçà duquel ils seront considérés comme inexistants. La valeur de ce seuil est déterminée de manière automatique par ajustements successifs de manière à ce que la composante connexe du graphe incluant l'arc de poids fort sélectionné comprenne entre 4 et 10 sommets (valeurs choisies empiriquement, qui nous permettent d'assurer la cohérence des noyaux extraits tout en rassemblant un nombre de mots suffisant pour le bon fonctionnement du traitement suivant). Les mots correspondant à ces sommets sont alors considérés comme formant une classe, et sont ôtés du graphe. On itère ensuite le processus de recherche décrit tant qu'il est possible par ce moyen d'extraire une nouvelle classe répondant aux critères requis.
On obtient ainsi de 40 à 45 [6] classes thématiques, restreintes mais ne comptant que des mots fortement associés à un thème.
Ceux-ci appartiennent tous à l'ensemble des mots retenus pour la classification lors de la première étape de construction de classes, et ont donc été initialement sélectionnés sur un critère de fréquence. Nous ne disposons en conséquence à ce stade que des mots à la fois fortement " colorés " thématiquement et apparaissant à de nombreuses reprises dans le corpus. Une dernière étape de calcul nous permet maintenant d'étendre les classes obtenues afin d'y adjoindre les mots-clefs de fréquence moindre. Les " prototypes de classes " dont nous disposons y sont mis à profit comme amorce d'un processus d'apprentissage de type supervisé, qui permet une bien meilleure précision de l'extraction.

5 Génération des classes définitives

Cette dernière partie du système de construction de classes de mots-clefs, dont l'objectif est d'accroître le nombre de termes rassemblés dans chaque classe, fonctionne selon un principe très simple : ajouter à chacune des classes produites précédemment les mots présentant sur les paragraphes qu'elle reconnaît une densité anormalement élevée relativement à leur fréquence moyenne sur le corpus – signe que ces mots possèdent un lien particulier avec le thème caractérisé par la classe de mots-clefs. Étant donnée la simplicité de la procédure mise en œuvre, nous pouvons maintenant travailler sur l'intégralité des 98 000 paragraphes du corpus, en considérant comme mots-clefs potentiels tous les noms et adjectifs y apparaissant plus de 100 fois, soit environ 3 600 mots.
À chacun des prototypes de classes $\mathcal{C}_i$ obtenus à l'étape précédente, on associe la liste LP_i des paragraphes qu'il reconnaît, puis l'on répète le traitement suivant :

- pour chaque classe $\mathcal{C}_i$, pour chaque mot-clef potentiel M_j, calculer

$$r_{ij} = \frac{\text{fréquence de } M_j \text{ sur } LP_i}{\text{fréquence de } M_j \text{ sur l'ensemble du corpus}}$$

- rechercher I et J tels que $r_{IJ} = max_{i,\,j}\,(r_{ij})$
- ajouter M_J à $\mathcal{C}_I$ et mettre à jour LP_I

tant que l'ajout du nouveau mot M_J permet d'augmenter le nombre moyen de mots de $\mathcal{C}_I$ par paragraphe de LP_I.

L'intérêt de la mise à jour des LP_i à mesure de la croissance des classes se comprend aisément en termes statistiques : lors des premières itérations de l'algorithme, les termes sélectionnés présentent des valeurs de r_{ij} très fortes (jusqu'à 200), ce qui compense l'amplitude de l'intervalle de confiance induit par la petite taille des LP_i (quelques dizaines de paragraphes). À mesure que la construction des classes de mots-clefs progresse, les valeurs de r_{ij} considérées sont de plus en plus faibles, mais cela est compensé par la réduction de l'amplitude de leur intervalle de confiance (due à la croissance des LP_i).
Le critère adopté pour décider du moment auquel interrompre la croissance d'une classe (diminution du nombre moyen de mots-clefs par paragraphe reconnu) est justifié par le fait qu'une baisse de cette densité moyenne correspond le plus souvent à l'ajout d'un mot tendant à étendre le thème original vers une notion annexe. Afin d'assurer la précision de la détection de thème réalisée grâce aux classes de mots-clefs, il est nécessaire qu'elles soient aussi précisément ciblées que possible, et cet assouplissement n'est donc pas souhaitable ; d'où le choix effectué de finaliser la classe considérée avant l'ajout de ce nouveau terme.

5. On constate expérimentalement que la qualité des résultats de l'algorithme présenté dans cette section se stabilise pour $n = 40$ environ.

6. Du fait du procédé de sélection aléatoire des paragraphes, les résultats peuvent varier légèrement d'une exécution du système à l'autre. C'est pourquoi nous ne donnons que des valeurs moyennes.

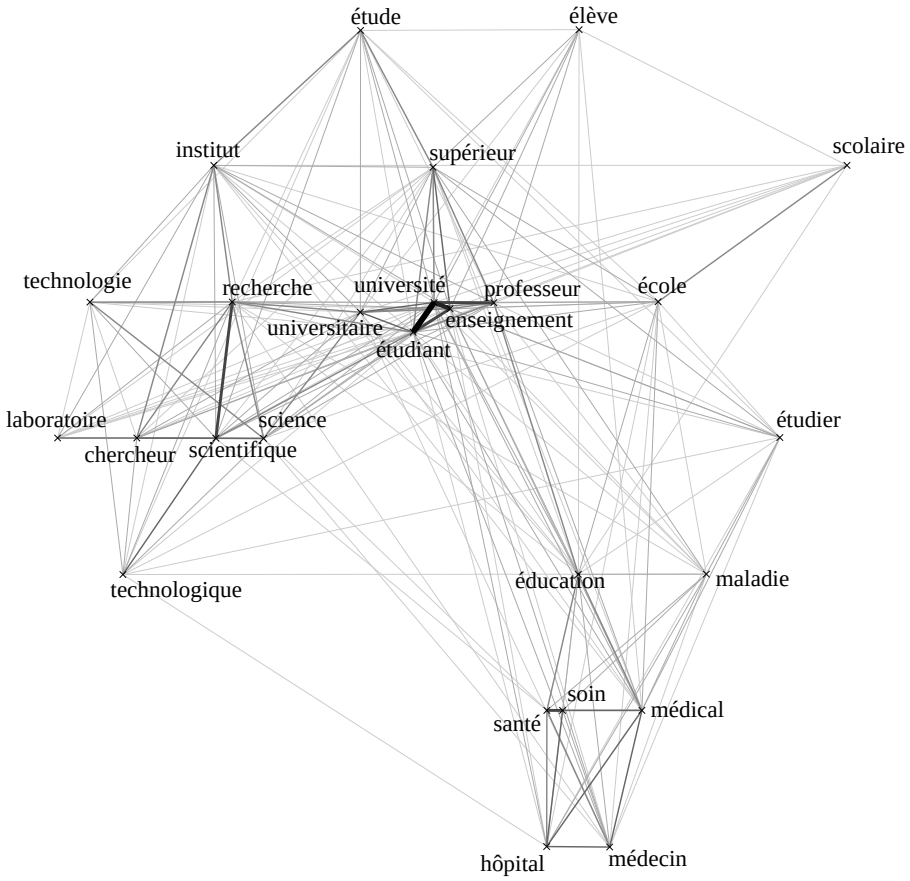

FIG. 2 – *Extrait du graphe de regroupement des mots-clefs potentiels construit pour une étude prenant en compte les 1 000 noms et adjectifs les plus fréquents du corpus, représenté pour un voisinage du mot " université ". Les positions des mots ont été calculées par un algorithme tentant de rapprocher d'autant plus deux sommets que l'arc les reliant a un poids important (arc épais sur la figure). Par souci de lisibilité, les arcs de poids 1 n'ont pas été représentés.*

Ce traitement est suivi d'un simple filtrage écartant les classes trop anecdotiques reconnaissant moins de 100 paragraphes, soit 1 ‰ du corpus – en général, une ou deux classes de mots-clefs entrent dans cette catégorie –, au terme duquel on dispose d'environ 40 classes de mots-clefs rassemblant 30 à 40 mots chacune.

Nommage des classes. Afin de favoriser l'exploitation subséquente de ces classes ainsi que leur compréhension par l'utilisateur, une heuristique très simple est employée pour associer à chaque classe un triplet de mots significatifs désignant sans ambiguïté le thème caractérisé par la classe de mots-clefs. Elle consiste à extraire de chaque classe de mots-clefs $\mathcal{C}$ un sous-ensemble de 3 mots tel que l'ensemble des paragraphes comprenant au moins l'un de ces mots couvre une partie la plus étendue possible de l'ensemble des paragraphes reconnus par $\mathcal{C}$. Cette contrainte est suffisante pour assurer que les 3 mots choisis sont assez fréquents, et ne correspondent pas à une composante absolument anecdotique du thème caractérisé par $\mathcal{C}$. De plus, afin de maximiser la " couverture " des paragraphes reconnus par $\mathcal{C}$, chacun des 3 mots sélectionnés est présent dans un nombre important de paragraphes où les 2 autres n'apparaissent pas, et le triplet fournit donc un indice de " l'extension " du thème sous-jacent à $\mathcal{C}$. Ainsi une classe dont la dénomination extraite est < enseignement / école / université > reconnaîtra-t-elle les paragraphes évoquant le système éducatif (*enseignement*) au sens large, du primaire (*école*) au supérieur (*université*).

6 Résultats

Nous évaluons dans cette section la qualité des résultats obtenus en fonction de deux critères : d'une part, la valeur descriptive des classes de mots-clefs et leur cohérence au regard d'une interprétation humaine (le volet " caractérisation de thèmes " de notre système), d'autre part leur efficacité en tant qu'outil de détection thématique.

6.1 Approche " intuitive "

Cette approche constitue la perception la plus directe de la qualité des classes de mots-clefs obtenues, ainsi que la plus subjective naturellement. Il nous a néanmoins semblé important, étant donné l'objectif fixé de caractérisation de

thèmes, de contrôler que les résultats se prêtent volontiers à une appréhension humaine. De ce point de vue, les classes de mots-clefs obtenues remplissent leur rôle et reflètent toutes clairement un thème précis, assez fidèlement décrit par le triplet de mots dont nous avons exposé le mode de sélection. Nous présentons ici le contenu de deux classes de mots-clefs obtenues par une exécution du système afin de permettre au lecteur de se faire une idée du type de résultats produits, ainsi que la liste des " noms " des 38 autres classes générées au cours de cette même exécution [7] :

- < enseignement / école / université > : { *alphabétisation, collège, diplôme, diplômé, école, éducatif, éducation, élève, enseignant, enseignement, enseigné, étudiant, étudié, instituteur, lycée, mathématique, primaire, pédagogique, scolaire, scolarisation, secondaire, universitaire, université* } ;
- < producteur / agriculture / céréale > : { *agriculteur, alimentation, aliment, approvisionnement, autosuffisance, blé, bétail, coton, céréale, céréalier, denrée, élevage, engrais, famine, fruit, grain, huile, intensif, irrigation, lait, légume, nourriture, oeuf, pain, pesticide, plantation, producteur, rendement, repas, riz, récolte, sac, saison, sec, semence, stock, sucre, surplus, sécheresse, vache, viande, vin, vivrier* } ;
- < armée / colonel / guérilla >, < informatique / satellite / machine >, < élection / voix / victoire >, < accord / barrière / douane >, < membre / réunion / comité >, < travail / emploi / salarié >, < minorité / langue / ethnique >, < règle / juridique / article >, < durée / professionnel / qualification >, < taux / baisse / croissance >, < maladie / hôpital / médical >, < alliance / pacte / traité >, < missile / conventionnel / arsenal >, < nation / conseil / international >, < territoire / occupation / colonisation >, < paix / conférence / frontalier >, < impôt / fiscal / assurance >, < pape / foi / évêque >, < réforme / ferme / propriétaire >, < technologie / ingénieur / biologique >, < crime / nazi / historien >, < prison / judiciaire / détenu >, < entreprise / investissement / infrastructure >, < quartier / parc / tourisme >, < campagne / démocrate / candidat >, < écrivain / roman / revue >, < télévision / image / journaliste >, < congrès / constitution / suprême >, < déficit / dette / budgétaire >, < roi / saoudite / chiite >, < forêt / pollution / sécheresse >, < unification / chancelier / est-allemand >, < fascisme / rhétorique / négation >, < asile / immigration / séjour >, < déchet / énergétique / centrale >, < film / cinéma / écran >, < financier / banque / crédit >, < théâtre / scène / musique >, < pétrolier / gaz / baril >.

Un lecteur du *Monde Diplomatique* pourra constater que les classes énumérées réalisent une couverture assez complète de la palette thématique développée par notre corpus. Il est particulièrement intéressant de noter les variations de granularité observables en fonction des domaines de prédilection de celui-ci : ainsi, si tous les " arts vivants " se trouvent rassemblés au sein de l'unique classe < théâtre / scène / musique >, on remarque un luxe de détails en ce qui concerne les questions économiques ou géopolitiques. Cette constatation montre à l'évidence l'intérêt d'une étude réalisée sur corpus sans présupposé sur la nature des résultats attendus.

Dans le même ordre d'idée, on peut observer l'émergence de thèmes liés à des faits particulièrement marquants de la période couverte par les archives étudiées, qui acquièrent de par le nombre de textes qui leur est consacré le statut de thèmes à part entière : ainsi la classe < territoire / occupation / colonisation > se réfère-t-elle au conflit israélo-palestinien, et < unification / chancelier / est-allemand > à l'unification de l'Allemagne. Deux classes révèlent également l'engagement du *Monde Diplomatique* dans les débats soulevés en France par les écrits d'auteurs révisionnistes comme Robert Faurisson (particulièrement actif durant la première moitié de la période considérée) ou Roger Garaudy (vers le milieu des années 90), d'une part par un retour sur les faits historiques (classe < crime / nazi / historien >) et, d'autre part, par une analyse de la rhétorique déployée par ces auteurs (< fascisme / rhétorique / négation >). On peut découvrir à la lecture des classes obtenues de nombreux autres indicateurs des choix " idéologiques " du *Monde Diplomatique*, telle la présence du mot *propagande* au sein de la classe < télévision / image / journaliste > ou celle de *lobby* dans < campagne / démocrate / candidat >, classe notamment centrée sur les campagnes électorales américaines.

Ainsi les classes obtenues permettent-elles, en identifiant les principaux thèmes du corpus et en jetant un premier éclairage sur leur mode de traitement, un véritable " survol " du contenu du corpus, ce qui présente un premier intérêt en soi, par exemple comme aide à l'exploration de grandes quantités de données textuelles.

Cette cohérence " intuitive " des classes de mots-clefs n'offre néanmoins aucune garantie concernant leur efficacité pour la tâche de détection de thèmes que nous visons. Nous nous attachons donc maintenant à évaluer leur pertinence de ce second point de vue, plus aisément quantifiable.

6.2 Efficacité des classes de mots-clefs pour la détection de thèmes

Cette section propose un ensemble d'indicateurs numériques permettant de juger la pertinence et la complétude de la détection de thèmes réalisée par nos classes de mots-clefs. Le critère d'évaluation considéré est donc cette fois plus objectif, bien qu'il laisse encore une certaine marge d'interprétation à l'expérimentateur chargé de décider si le thème détecté dans un paragraphe y est réellement présent.

Avant de présenter et analyser les résultats de cette validation, nous en précisons ici les conditions :

- outre le principe de cooccurrence de deux mots adopté jusqu'ici, nous avons été amenés à développer d'autres

7. L'intégralité des classes de mots-clefs obtenues est disponible sur demande auprès des auteurs.

critères de détection de thèmes afin, notamment, d'améliorer la précision de celle-ci, critères que nous présentons dans un premier temps ;
– nous décrivons ensuite les différents indices employés pour évaluer la qualité de la détection de thèmes réalisée, ainsi que les procédures expérimentales mises en œuvre afin de les calculer.

Critères de détection élaborés. Une variation élémentaire du premier critère de détection de thème par cooccurrence de deux mots-clefs dans un paragraphe consiste à accroître la sélectivité de celui-ci en augmentant le nombre de mots-clefs requis à trois. Si l'on obtient ainsi, à l'évidence, une fiabilité plus grande de la détection, il est probable que cela soit au prix d'une baisse importante de la mesure de rappel de la détection.
Un troisième critère tentant de concilier précision et rappel a été mis au point en faisant usage d'un niveau de structuration de notre corpus journalistique non exploité jusqu'ici : son découpage en articles. En effet, si une classe de mots-clefs reconnaît plusieurs paragraphes au sein d'un même article, il est vraisemblable que le thème caractérisé par cette classe relève de la problématique abordée par l'auteur dans ce texte, et la probabilité de le voir apparaître ailleurs dans l'article s'en trouve accrue. Inversement, il est assez peu courant de voir un thème apparaître de manière isolée dans un unique paragraphe d'un article – et si cela se produit, le vocabulaire spécifique à ce thème se trouvera concentré dans le paragraphe en question, augmentant ainsi le nombre de mots-clefs potentiels. Pour refléter ce double principe, on souhaite rendre le critère de détection d'un thème donné plus sensible si ce thème apparaît déjà dans l'article étudié, et plus strict si ce n'est pas le cas. Un nouveau critère est donc défini de la manière suivante : partant d'une détection de thèmes par le critère de co-présence de trois mots-clefs présenté ci-dessus, on étudie pour chaque article le nombre de paragraphes détectés comme abordant un thème donné.

– Si ce nombre est supérieur ou égal à deux, alors le seuil de détection pour ce thème dans l'article considéré est baissé à deux mots-clefs, permettant potentiellement la détection du thème dans un plus grand nombre de paragraphes de l'article,
– sinon, ce seuil est réhaussé à quatre et le critère devient donc plus strict, remettant en cause la détection du thème dans le pararaphe isolé.

Nous présentons maintenant la procédure de validation employée afin de juger de la qualité des détections réalisées par ces trois critères.

Indices qualitatifs. Trois indices numériques sont pris en compte afin de juger du succès de la détection de thèmes réalisée par les classes de mots-clefs construites :

– la " couverture " du corpus, c'est-à-dire la proportion de ses paragraphes auxquels il a été possible d'attribuer au moins un thème. Cette valeur est également exprimée en pourcentage des mots du corpus rassemblés par ces paragraphes ;
– la précision de la reconnaissance de thèmes, qui indique le pourcentage global de détections pertinentes parmi celles réalisées par les classes produites. Cette valeur est calculée sur une sélection aléatoire de 1 000 paragraphes auxquels au moins un thème est attribué, en contrôlant manuellement pour chacun d'eux la présence du (des) thème(s) détecté(s), sans aucune exigence concernant l'importance du thème à l'échelle du paragraphe ;
– le rappel observé, enfin, n'a été calculé que pour le thème " éducation " en recherchant manuellement dans le corpus 100 paragraphes l'abordant en effet, et en contrôlant quelle proportion de ceux-ci est reconnue par la classe < enseignement / école / université >.

Le tableau 1 rassemble l'ensemble des valeurs de ces indices, calculées pour chacun des critères de détection présentés.

Éléments d'analyse. Une première constatation possible à la lecture de ce tableau est la différence assez importante existant entre les valeurs de la couverture du corpus exprimées en paragraphes et en nombre de mots. Cette divergence s'explique par le fait que plus un paragraphe contient de mots, meilleure est la prise qu'il donne à une détection de thème par cooccurrence de mots-clefs. Ainsi, pour le troisième critère, on calcule que les paragraphes reconnus comme abordant au moins un thème sont en moyenne 35 % plus longs que les paragraphes sans thème détecté.
Le rappel mentionné, calculé pour un seul thème, n'a à l'évidence qu'une valeur indicative et est susceptible de varier d'un thème à l'autre ; il permet néanmoins de constater que le dernier critère décrit, qui prend en compte la structuration en articles du corpus, parvient comme souhaité à concilier précision et rappel de la détection : on n'observe aucune perte de précision par rapport au critère de cooccurrence de trois mots-clefs, et une faible baisse du rappel par rapport au critère ne faisant usage que de deux mots.

7 Conclusion

Le système présenté dans cet article permet l'acquisition à partir de données textuelles, sans faire usage de connaissances complémentaires ni requérir d'intervention humaine, d'un ensemble de classes de mots-clefs caractérisant de manière complète le contenu thématique du corpus étudié, et fournissant en outre des indications concernant la manière dont les divers thèmes y sont traités. Dans ce sens, il constitue donc un outil intéressant pour l'analyse de grandes quantités de textes.
Les classes obtenues permettent en outre de détecter avec une relativement bonne précision les occurrences des thèmes du corpus dans ses paragraphes. Cette détection peut être directement exploitée pour de nombreuses tâches dans le domaine du TAL : nous en faisons ainsi usage afin de découper un corpus en plusieurs sous-corpus thématiques ho-

	Couverture		Précision	Rappel " éducation "
	paragraphes	mots	(paragraphes)	(paragraphes)
Deux mots	66 %	70 %	55 %	65 %
Trois mots	32 %	40 %	85 %	35 %
Trois mots + structure	58 %	65 %	85 %	63 %

TAB. 1 – *Couverture, précision et rappel de la détection de thèmes réalisée, calculés pour les trois critères employés. Le critère désigné par " trois mots + structure " est le dernier de ceux présentés dans la section 6.2, qui fait usage de la structuration en articles du corpus.*

mogènes, afin de permettre une désambiguïsation rapide des mots polysémiques dans une optique de construction automatique de lexiques sémantiques.

La qualité de cette détection est néanmoins limitée par son rappel, et notamment par le fait qu'un tiers environ des paragraphes étudiés n'est pas caractérisé. La taille et le nombre relativement importants des classes de mots-clefs obtenues donnent à penser que le facteur limitant est ici le mode de détection de thèmes retenu, consistant à faire usage de mots-clefs, plus que les performances du système de construction de ces classes. On constate en effet que de nombreux paragraphes abordent un thème de manière implicite, sans faire usage d'un vocabulaire marqué, parce que leur environnement textuel (paragraphe précédent, titre de l'article...) indique déjà clairement ce thème au lecteur.

Il serait donc probablement profitable de retenir le principe d'amorçage d'un apprentissage supervisé par une classification statistique présenté en section 5, en remplaçant l'algorithme assez naïf employé pour générer les classes définitives par l'un de ceux décrits dans [4] ou [3], afin d'obtenir un modèle de langage permettant une détection plus fine des occurrences de thèmes.

Remerciements

Les auteurs tiennent à remercier tout particulièrement I.-C. Lerman pour son assistance au cours de cette recherche, et notamment pour le travail réalisé sur l'adaptation de CHAVL à leurs besoins spécifiques.

Références

[1] Susan Armstrong. Multext : Multilingual Text Tools and Corpora. In H. Feldweg and W. Hinrichs, editors, *Lexikon und Text*, pages 107–119, 1996.

[2] Susan Armstrong, Pierrette Bouillon, and Gilbert Robert. Tagger Overview. Rapport de recherche, ISSCO, 1995.

[3] B. Bigi, R. De Mori, M. El-Beze, and T. Spriet. Combined Models for Topic Spotting and Topic-dependent Language Modeling. In S. Furui, B.H. Huang, and Wu Chu, editors, *1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 535–542, 1997.

[4] Armelle Brun. *Détection de thème et adaptation des modèles de langage pour la reconnaissance automatique de la parole*. PhD thesis, Université Henri Poincaré, Nancy I, 2003.

[5] Olivier Ferret and Brigite Grau. Utiliser des corpus pour amorcer une analyse thématique. *TAL (traitement automatique des langues), numéro spécial Traitement automatique des langues et linguistique de corpus*, 42(2), 2001.

[6] Marti A. Hearst. Multi-Paragraph Segmentation of Expository Texts. In *ACL'94 (32th Annual Meeting of the Association for Computational Linguistics)*, Las Cruces, NM, USA, 1994.

[7] Roland Kuhn and Renato de Mori. A Cache-Based Natural Language Model for Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583, 1990.

[8] Israël-César Lerman. Foundations in the Likelihood Linkage Analysis Classification Method. *Applied Stochastic Models and Data Analysis*, 7:69–76, 1991.

[9] Israël-César Lerman, Henri Leredde, and Philippe Peter. Principes et calculs de la méthode implantée dans le programme CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance du Lien) – deuxième partie. *Revue de* MODULAD, 13:63–90, 1994.

[10] Diane J. Litman and Rebecca J. Passonneau. Combining Multiple Knowledge Sources for Discourse Segmentation. In *ACL'95 (33th Annual Meeting of the Association for Computational Linguistics)*, Montréal, Québec, Canada, 1995.

[11] Ronan Pichon and Pascale Sébillot. From Corpus to Lexicon: from Contexts to Semantic Features. In Barbara Lewandowska-Tomaszczyk and Patrick James Melia, editors, *PALC'99 (Practical Applications in Language Corpora), Lodz studies in language*, volume 1. Peter Lang, 2000.

[12] Mathias Rossignol and Pascale Sébillot. Automatic Generation of Sets of Keywords for Theme Detection and Characterization. In Annie Morin and Pascale Sébillot, editors, *JADT 2002 (Journées internationales d'Analyse des Données Textuelles)*, Saint-Malo, France, 2002.

[13] Gerard Salton, Amit Singhal, Chris Buckley, and Mandar Mitra. Automatic Text Decomposition Using Text Segments and Text Themes. In *UK Conference on Hypertext*, pages 53–65, 1996.