Trois approches du GREYC pour la classification de textes

Thierry Charnois Antoine Doucet Yann Mathet François Rioult
GREYC, Université de Caen, CNRS UMR 6072
Bd Maréchal Juin, BP 5186, 14032 Caen Cedex
{charnois, doucet, mathet, frioult}@info.unicaen.fr

Résumé. Cet article présente la participation de l'équipe du GREYC à DEFT'08, en détaillant les différentes approches mises en place ainsi que les résultats obtenus. Plusieurs techniques très différentes ont été étudiées et mises en oeuvre. D'une part, un traitement à base de n-grammes a constitué un classifieur indépendant. D'autre part, deux autres traitements s'appuient sur un classifieur supervisé par règles d'association, qu'ils alimentent chacun avec des indices provenant d'une chaîne de traitements linguistiques pour l'un, et d'extraction de séquences pour l'autre.

Abstract. In this paper, we present the various approaches and corresponding results of the GREYC laboratory to the DEFT'08 challenge. A couple of distinct techniques were experimented with. On the one hand, an n-gram based classifier was tested. On the other hand, two different types of data were fed to a supervised classifier, based on association rules : 1) linguistic markers, and 2) discontinuous word sequences.

Mots-clés : Fouille de texte, classification par n-grammes, classification par règles d'association, TALN, séquences de mots.

Keywords: Text data mining, n-gram classification, association rule-based classification, NLP, word sequences.

1 Introduction

Le thème de cette édition 2008 de DEFT concerne la classification de textes en catégorie et en genre. Pour l'entraînement de la tâche, deux corpus sont disponibles. Le premier est un ensemble d'articles du journal Le Monde et d'articles de Wikipédia avec un double étiquetage : l'un en genre (Le Monde ou Wikipédia) et l'autre en catégorie thématique (sport, art, économie, *etc.*). Le deuxième corpus est également un ensemble d'articles issus des mêmes sources, mais avec un simple étiquetage en catégorie (les catégories de ce corpus étant différentes du premier). Pour le test, deux corpus sans étiquette ont été fournis et la tâche consiste à reconnaître le genre et la catégorie du premier corpus, la catégorie pour le deuxième.

Le laboratoire GREYC présente une équipe à DEFT pour la troisième année consécutive. Pour cette édition 2008, sa composition ainsi que les techniques mises en oeuvre sont en partie issues de la session précédente, et en partie nouvelles :

– Une approche complète à base de n-grammes, conçue et utilisée à l'occasion de DEFT'07, a été reprise et adaptée aux spécificités de DEFT'08,

– Un classifieur supervisé par règles d'association, utilisé depuis DEFT'06, est chaque année « alimenté » par différentes chaînes de traitement adaptées aux tâches spécifiques de chaque défi.

Pour cette session 2008, une chaîne de traitements linguistiques basée sur la plate-forme Linguastream a été conçue pour créer des indices concernant la classification en genres du corpus 1, et un autre traitement par extraction de séquences répondant ainsi à la classification en catégories des corpus 1 et 2.

Nous présentons dans les parties qui suivent chacune des trois approches, et terminerons pas une analyse comparative de nos résultats avant de conclure.

2 Un classifieur à base de n-grammes

Un classifieur autonome à base de n-grammes a été développé au GREYC en 2007 à l'occasion du défi DEFT'07, il est donc utilisé pour la deuxième année consécutive. Il a été conçu nativement pour gérer autant de catégories qu'on le souhaite, et a pu être adapté relativement facilement à la session DEFT'08 dans ses grands principes, même si un certain nombre d'aménagements techniques ont dû être effectués. Le présent exposé étant concis, nous invitons le lecteur désireux de connaître le détail de ce traitement à se référer aux actes de DEFT'07 (Vernier *et al.*, 2007).

2.1 Principe

La technique des n-grammes consiste à observer les collocations contiguës sur une fenêtre de n tokens consécutifs d'un flux, et à essayer de tirer de ces observations des régularités relatives à un aspect particulier de ce flux (Stubbs & Barth, 2003). Par exemple, certains n-grammes seront caractéristiques de tel type de corpus car très récurrents dans ce dernier, et beaucoup plus rares ailleurs. Pour illustrer de façon très simplifiée l'hypothèse de cette approche dans DEFT'08, nous espérons trouver des n-grammes très caractéristiques d'une catégorie par rapport aux autres catégories. Par exemple « en troisième division » ou « un match difficile » sont plutôt caractéristiques de la catégorie SPORT, tandis que « présentateur du JT », « une émission débridée » seront quant à eux plutôt caractéristiques de la catégorie TELEVISION. Ainsi, lors de l'analyse d'un texte du corpus de test, si l'on tombe sur le tri-gramme « présentateur du JT », nous serons tentés de ranger ce texte en catégorie TELEVISION. Bien sûr, deux difficultés apparaissent : il se peut qu'au cours de l'apprentissage, un même n-gramme soit présent dans des textes issus de différentes catégories ; et il est fréquent que lors du test, nous trouvions au sein d'un même texte des n-grammes issus de différentes catégories, rendant le choix plus difficile. L'idée que nous mettons en œuvre pour pallier ces difficultés est de deux ordres : ne retenir pour chaque catégorie que les n-grammes les plus discriminants, c'est-à-dire ceux étant le moins susceptibles d'apparaître dans des textes appartenant à d'autres catégories ; pondérer les n-grammes, c'est-à-dire associer à chacun un poids d'autant plus important qu'il apparaît fréquemment dans sa catégorie cible relativement aux autres catégories. Puis, pour chacune des catégories, sommer les poids de tous les n-grammes (de cette catégorie) trouvés dans le texte testé. De la sorte, nous obtenons une note globale pour chaque catégorie, que nous pouvons mettre en balance avec les notes globales obtenues pour les autres catégories.

2.2 Apprentissage

La collecte des n-grammes est effectuée pour chaque catégorie (ou genre). Nous extrayons dans ce but la collection des n-grammes dans chaque catégorie, puis calculons les n-grammes émergents par comparaison des n-grammes d'une catégorie à ceux des autres catégories :
– si un n-gramme est absent des autres catégories, on lui attribue un poids infini ;
– s'il est présent dans les autres catégories, on lui associe un poids égal au rapport entre sa fréquence relative dans la catégorie à caractériser et sa fréquence dans les autres.
Le n-gramme sera finalement utilisé dans le classifieur si son poids est supérieur à un certain seuil paramétrable.

2.3 Classification : choisir un genre ou une catégorie

Assigner une catégorie à un texte du corpus de test consiste à parcourir le texte au moyen d'une fenêtre de longueur n, et à recommander une catégorie pour laquelle un n-gramme découvert est émergent. Nous obtenons pour le texte de test autant de sommes de poids qu'il y a de catégories. Ces sommes correspondent à l'indice de confiance que l'on peut accorder à chaque catégorie. Nous pouvons alors assigner comme catégorie celle obtenant la somme de poids la plus élevée. Par ailleurs, notre application permet de combiner différents traitements à l'envi, par exemple bi-grammes et tri-grammes, selon que le corpus se prête mieux à telle ou telle combinaison.

2.4 Application aux différentes tâches

2.4.1 Méthode générale et ajustements

Chacun des deux corpus d'apprentissage fournis a tout d'abord été coupé en deux parties égales, constituant pour notre phase d'apprentissage un corpus d'apprentissage (par ex. la première moitié) et un corpus de test (par ex. la seconde moitié). De la sorte, nous avons une vision non biaisée des performances des traitements effectués, ce qui est indispensable pour réaliser les paramétrages les plus adéquats. Les apprentissages et tests ainsi réalisés, nous avons observé les résultats et constaté que certaines catégories étaient sous ou sur représentées, selon les rapports entre les rappels obtenus par chacune. C'est alors que nous appliquons des coefficients d'ajustement, afin que d'une part une catégorie sous représentée soit privilégiée (il suffit pour cela de multiplier les valeurs fournies par leurs n-grammes par un coefficient supérieur à 1), et d'autre part qu'une catégorie sur-représentée soit défavorisée (application d'un coefficient inférieur à 1). À l'issue de ces réglages, réalisés empiriquement par essais successifs (une automatisation nous aurait permis d'affiner nettement le réglage obtenu), nous obtenons une répartition en catégories beaucoup plus homogène en terme de rappel, et comme escompté, une amélioration du F-score.

Utilisée pour l'exécution 1 du défi, cette méthode donne un F-score de 96.4% sur le genre, 84.9% (tâche 1) et 83.7% (tâche 2) pour les catégories.

2.4.2 Test d'une hypothèse de DEFT'08 : la connaissance du genre aide-t-elle à mieux trouver la catégorie ?

La partie « GENRE » de la tâche 1 donnant lieu à des résultats extrêmement élevés (supérieurs à 96%), il est tentant d'essayer d'en tirer parti pour améliorer les scores sur la tâche 2. Le principe testé est le suivant :

1. appliquer le traitement en GENRE au corpus 2. On trouve alors pour chacun des textes, avec un degré de confiance supérieur à 96%, le genre de ce dernier.
2. diviser ainsi le corpus d'apprentissage en deux sous-corpus, l'un contenant les textes jugés appartenir au genre « Wikipedia », l'autre ceux jugés appartenir au genre « Le Monde ».
3. appliquer alors, de façon séparée, l'apprentissage en n-grammes sur les deux sous-corpus.
4. procéder de la même façon pour la phase de test.

Les résultats sont finalement légèrement inférieurs au traitement classique, ce qui ne permet pas de répondre positivement à la question posée. En fait, la réponse doit être plus nuancée :

1. le fait de particulariser le traitement selon le genre donne vraisemblablement des n-grammes plus précis que dans le cas général.
2. mais ceci est contrecarré par le fait que les deux demi-corpus d'apprentissage correspondant sont chacun, évidemment, deux fois plus petits que le corpus originel.

Il est donc probable qu'avec des corpus de plus en plus grands, compte tenu du fait que le gain obtenu par une augmentation de la taille des corpus est forcément limité asymptotiquement, nous finirions par effacer l'inconvénient mentionné en 2), et mettre en avant de façon enfin positive le gain obtenu en 1).

3 Approche TALN

L'approche TALN s'est focalisée sur la classification en genre (corpus 1). Elle vise à combiner analyse linguistique et apprentissage automatique. Nous reprenons, en les adaptant à la tâche fixée, les principes généraux déjà présentés lors des éditions 2006 et 2007 du défi DEFT (Widlöcher *et al.*, 2006) et (Vernier *et al.*, 2007), principalement :
– une phase de modélisation pour dégager des critères linguistiques génériques et pertinents pour la classification ;
– la réalisation d'une chaîne de traitements pour repérer ces indices et les marquer dans les corpus ;
– enfin, la mise au point d'un classifieur à partir du marquage textuel des indices (détaillé en section 5).

3.1 Modélisation linguistique

Deux genres – deux styles

Ce travail s'appuie sur une observation minutieuse du corpus et de sa nature pour procéder à une catégorisation linguistique des deux genres à discriminer. Elle repose sur l'hypothèse sous-jacente suivante : les deux genres sont révélateurs de deux styles, l'un journalistique (Le

Monde), l'autre encyclopédique (Wikipedia), qui vont utiliser leurs propres marques linguistiques. Le premier est plus expositif ou narratif, incluant des citations, et induit l'usage d'une palette assez large de formes langagières (temps verbaux variés, interrogation, négation, citation...).

À l'opposé, les textes de Wikipédia, par nature encyclopédique, nous semblent relever du style définitoire plus spécifique, que souligne l'usage fréquent de marques méta-linguistiques comme « être un », « désigne », « définit », *etc*. (*cf*. (Chaurand & Mazière, 1990) sur la notion d'acte définitoire).

Lors du traitement de ces marques, seules celles apparaissant en tête du texte, c'est-à-dire **en première phrase**, ont été considérées. Cette contrainte répond à l'hypothèse selon laquelle cette position joue un rôle priviligié dans l'organisation discursive et en particulier nous pensons que les marques discriminantes en matière de genre textuel sont celles qui sont situées dans cette position.

Nous passons maintenant en revue les différents indices retenus comme propriétés discriminantes et caractéristiques des deux genres.

Indices Wikipédia Ce type d'indice est en nombre restreint. Il concerne les verbes induisant un mode définitoire : forme verbale « être » suivie d'un déterminant, « désigner », « définir », « signifier », etc. Nous y avons ajouté les marques exprimant une naissance ou un décès : « né le »... Au moins l'un des deux indices apparaît[1] dans 6107 articles de Wikipédia contre 595 articles du Monde.

Indices pour Le Monde Les indices que nous considérons comme caractéristiques de ce genre sont relativement moins nombreux en terme d'occurrences, mais plus divers :
- les formes énonciatives (pronoms personnels des 1ère et 2ème personne, marques de citation) indiquant la présence d'un locuteur, cas typique de l'interview. L'une des formes est présente dans 1957 articles du Monde (respectivement 148 articles de Wikipédia) ;
- les temps verbaux (passé, futur, conditionnel présent) significatifs de la narration (versus le présent atemporel de la définition) : 4928 articles du Monde (resp 1035) ;
- les formes marquant une question, une exclamation ou une négation sont présentes dans 1276 articles du Monde (resp 158) ;
- les marques anaphoriques 1223 articles du Monde (resp 301) ;
- les formes de type « c'est un » et les formes impersonnelles (« il y a », « il est difficile de », *etc*.) : 687 (resp 126).

3.2 Réalisation informatique

Le repérage et le marquage des indices linguistiques a été réalisé à l'aide de la plate-forme LinguaStram (Bilhaut & Widlöcher, 2006) dédiée au TALN. Une chaîne de traitements séquentiels a été mise au point. Un premier traitement extrait la première phrase de chaque texte du corpus. Puis un composant morpho-syntaxique se charge de donner pour chaque mot sa catégorie syntaxique et sa forme lemmatisée[2]. Le coeur du traitement se compose de grammaires DCG (une

[1] Le calcul a été opéré sur le corpus d'apprentissage et sur la première phrase de chaque texte.
[2] nous utilisons ici l'outil bien connu TreeTagger

par indice) qui opèrent le marquage des indices linguistiques recherchés. En bout de chaîne, un dernier composant produit une matrice dans laquelle chaque ligne correspond à un texte du corpus et chaque colonne à un attribut étiqueté par un indice. La valeur de cet indice est le nombre d'occurrences de l'objet ou 0 si l'indice est absent de la première phrase.

3.3 Classification

Les règles de classification sont produites automatiquement à partir de la matrice (voir section 5). Le traitement sur la première phrase donne un F-score de 86%. On observe cependant dans la matrice un nombre élevé de lignes contenant une forte proportion de valeurs nulles (par exemple 15% des lignes n'ont qu'un attribut non nul). La prise en compte du texte dans sa totalité mérite d'être expérimentée. Cela nécessite de distinguer les critères à n'appliquer que sur la tête du texte d'une part, et d'autre part ceux à appliquer sur tout le texte. Par exemple, la forme « être + déterminant » a, comme on l'a vu, un nombre d'occurrences très faible au sein de la première phrase du Monde (relativement aux textes de Wikipédia). La probabilité d'apparition de cette forme est sans doute beaucoup plus importante dans les phrases suivantes ; le marquage de cet indice dans tout le texte ferait donc perdre à ce critère son pouvoir discrimant.

4 Utilisation de séquences discontinues de mots

Motivés par de précédentes expériences en recherche d'information (Doucet, 2004), nous avons voulu tester l'efficacité de l'utilisation de descripteurs séquentiels dans le cadre de la classification textuelle supervisée.

4.1 Motivation

La majorité des méthodes d'exploitation de contenus textuels adoptent un modèle de type « sac de mots », considérant implicitement les occurrences de mots comme des faits indépendants.

Si ces méthodes permettent d'obtenir de bons résultats, il semble toutefois raisonnable de penser que l'intégration d'une information supplémentaire, telle que la prise en compte des unités lexicales complexes, doit permettre d'améliorer la performance des systèmes de classification.

L'un des défauts d'une telle représentation documentaire est qu'elle ne tient pas compte des positions relatives des mots dans le document, ce qui semble intuitivement anormal, étant donné que des mots sont plus probablement liés s'ils apparaissent côte à côte plutôt qu'au début et à la fin d'un livre. En outre, l'occurrence simultanée de plusieurs mots induit souvent un sens différent de « l'addition » de la signification de ces mots pris individuellement. Par exemple, si l'on dit d'une personne qu'elle a « la main verte », on ne parle pas réellement de la couleur de sa main ; cela signifie que cette personne est douée pour le jardinage. Les expressions métaphoriques sont source de nombreux exemples de ce type (par exemple « avoir un chat dans la gorge »).

De nombreuses méthodes existent pour extraire des unités lexicales complexes. Elles reposent sur des critères statistiques, sur des critères syntaxiques, ou bien encore sur ces deux types de critères à la fois (application de méthodes statistiques après un filtrage syntaxique, par exemple).

Un défaut des méthodes statistiques pures est que, pour des raisons de complexité calculatoire, il est impossible en pratique de calculer une mesure d'association pour tous les ensembles de mots pouvant éventuellement former une unité lexicale.

Ainsi, les chercheurs ont toujours placé un certain nombre de restrictions lors de l'extraction d'unités multi-mots, en n'appliquant de mesures d'association qu'aux ensembles de mots respectant certaines contraintes, comme par exemple une longueur maximale, ou des positions d'occurrence rigides (positions relatives fixes ou restreintes par un nombre maximum de mots autorisés entre deux mots d'une unité lexicale).

La restriction la plus courante est d'imposer l'adjacence des termes (n-grammes), ce qui implique une perte d'information considérable. Par exemple si le mot « et » intervient entre deux autres mots, ils sont très certainement liés mais cela ne peut être pris en considération.

4.2 Séquences Fréquentes Maximales

Pour permettre l'extraction de séquences de mots sans restriction sur la longueur des séquences, ni sur la distance séparant leurs composants, nous proposons l'utilisation de séquences fréquentes maximales (SFM) (Ahonen-Myka & Doucet, 2005)).

Définition. Une séquence est dite fréquente si elle apparaît dans un nombre de phrases supérieur à un seuil de fréquence donné. Elle est maximale si on ne peut y insérer aucun autre mot sans pour autant faire descendre sa fréquence sous ce seuil.

Exemple. L'utilisation des SFMs permet d'appréhender le fait que la séquence « président Bush » apparaît dans chacun des 2 fragments textuels suivants, ce qui ne serait notamment pas le cas avec une méthode nécessitant des contraintes d'adjacence :

```
...Le président des Etats Unis George Bush...
...Président George W. Bush...
```

4.3 Application à DEFT'08

Apprentissage. À l'aide de l'étiquetage fourni dans les données d'apprentissage, nous avons construit une (sous-)collection de documents correspondant à chaque genre et chaque catégorie. Nous avons alors lancé l'extraction des SFMs dans chacune de ces collections, obtenant ainsi un ensemble de SFMs représentatives du genre et/ou de la catégorie correspondante. Afin de faciliter les comparaisons entre SFMs, nous avons finalement décomposé chaque SFM en chacune des paires de mots qui les composent. Nous avons alors utilisé le classifieur présenté en Section 5 afin d'extraire les règles à appliquer au corpus de test.

Test. Afin d'extraire nos séquences dans des proportions comparables, nous avons formé plusieurs sous-collections disjointes à l'aide de l'algorithme de clustering $k-means$. L'extraction des SFMs a alors été conduite parallèllement dans chacune des sous-collections. Cette approche

« diviser pour mieux régner » permet une extraction de séquences plus rapide et plus exhaustive (Doucet & Ahonen-Myka, 2006). Après normalisation, chaque document du corpus de test était associé à un ensemble de paires ordonnées issues de l'apprentissage. Les règles du classifieur fournissent la décision finale.

5 Classification surpervisée

Les données étiquettées par les approches TALN et extraction de séquences décrivent les textes du corpus à l'aide de descripteurs. Un descripteur particulier désigne l'appartenance à la classe. Pour répondre au défi, nous avons calculé sur ces données supervisées un classifieur à base de règles d'association, entraîné par 10-cross-validation.

5.1 Classification surpervisée à base d'association

Une règle de classification est une règle d'association (Agrawal *et al.*, 1996) concluant sur un attribut de classe. Si de telles règles peuvent être découvertes dans les données, l'intuition indique qu'elles peuvent aider à classer les textes supportant les descripteurs de la prémisse de la règle. CMAR (Li *et al.*, 2001) (Classification based on Multiple class-Association Rules) est une méthode populaire de classification à base de règles d'association. Les règles sont mesurées par un indice de corrélation fourni par un χ^2 normalisé. La redondance des règles est évitée en ne conservant que celles qui sont à prémisse minimale. Un nouvel exemple sera classé à l'issue d'un vote réalisé par toutes les règles qui s'appliquent, selon leur pondération.

Pour le défi, nous avons utilisé notre adaptation de CMAR, qui consiste à généraliser la forme des règles de classification en autorisant la présence de négation en prémisse (pour caractériser des objets contenant un motif mais en excluant un autre) ou en conclusion (pour des objets excluant une classe) (Rioult *et al.*, 2008).

5.2 Application aux données du défi

Les données de la méthode TALN (*cf.* section 3) ont été utilisées pour effectuer la classification en genre (Le Monde / Wikipedia) et les séquences (*cf.* section 4) ont classé les textes en catégories.

5.2.1 Données TALN - genre

Quelques expériences rapides ont permis de constater que la seule règle `être + déterminant` $\rightarrow$ `Wikipedia`, de confiance 73% (présente 4703 fois dans les 6398 textes de classe Wikipedia et uniquement 446 fois dans les 88825 textes du Monde), permet d'obtenir sur l'échantillon d'apprentissage un F-score moyen de 85.2%. Le F-score théorique d'un classifieur utilisant uniquement cette règle est de 86%, qui est le résultat obtenu sur les données de test (exécution 2).

Nous avons tenté d'utiliser des règles moins fréquentes, mais les performances étaient moins bonnes. Même en expérimentant des règles justifiées d'un point de vue de la sémantique, nous n'avons pu améliorer les performances de ce simple indice.

5.2.2 Données séquences - catégorie

Les données de séquences fréquentes extraites répertorient 11156 paires de mots et sont très volumineuses. Après un filtrage des plus fréquentes (présentes dans plus de 500 textes), nous obtenons une matrice de 15215 textes et 10414 descripteurs.

Les méthodes d'extraction de connaissance à base de motifs, telles que les règles d'associations utilisées ici pour calculer un classifieur, sont très coûteuses en temps de calcul. Dans le pire des cas, elles sont polynomiales en nombre d'objets et exponentielles en nombre d'attributs. La taille de la matrice à traiter ici est très pénalisante car elle contient beaucoup d'attributs.

Afin d'effectuer les calculs dans le temps imparti, nous avons dû restreindre les règles à des prémisses contenant une unique séquence. Malgré cette simplification, le F-score vérifié par 10-cross-validation était très bon : entre 92 et 93%.

Hélas, les séquences ont été obtenues sur l'intégralité de l'échantillon d'apprentissage, et nos expérimentations ont été fortement biaisées. D'ailleurs, les performances obtenues lors du défi sont très médiocres (67% pour la tâche 1, et 32% pour la tâche 2). Ces résultats ne remettent cependant pas en cause la qualité des séquences calculées ni la méthode de décision utilisée. Si le calcul des séquences avait été intégré au processus de validation croisée, le classifieur aurait révélé de nettement moins bonnes performances et nous aurions alors cherché à les améliorer. Nous proposons, lors de la réunion des participants, de donner des scores plus représentatifs du potentiel de notre méthode.

6 Analyse - perspectives

La table 1 indique les F-scores obtenus à l'aide de nos différentes méthodes. L'exécution 1 correspond à la méthode n-grammes (section 2), l'exécution 2 correspond à la méthode TALN et extraction de séquences, puis décision par classification par règles d'association (sections 3 à 5). L'exécution 3 utilise la méthode n-grammes avec un paramétrage différent pour les tâches 1 genre et 2 catégorie, et la méthode séquences pour les catégories de la tâche 1.

	exécution 1 (n-grammes)	exécution 2 TALN + séquences	exécution 3 mix 3 méthodes	moyenne participants
tâche 1 genre	96.4	85.6	96.4	95.9
tâche 1 catégorie	84.9	67.2	67.2	82.6
tâche 2 catégorie	83.7	32.8	81.5	81.1

TAB. 1 – Récapitulatif des F-scores (en %) obtenus avec les différentes méthodes.

Une analyse des résultats montre que la méthode n-grammes donne les meilleurs résultats. Ces résultats se situent dans la moyenne de l'ensemble des participants à DEFT. Par ailleurs, ils restent homogènes sur les trois tâches, contrairement à l'an dernier. La méthode des n-grammes tire ainsi parti de la taille importante des corpus nécessaire à l'établissement de bonnes performances. En revanche, la connaissance du genre ne permet pas d'améliorer les résultats pour les raisons évoquées en section 2.4.2. Enfin, il est à noter que cette approche est peu coûteuse en ressources machine et ne demande que quelques minutes pour l'apprentissage et le test.

Les résultats obtenus par la méthode TALN sont honorables et valident l'intérêt de cette approche tant pour confirmer des hypothèses linguistiques, que pour son originalité. Originalité

qui tient à son aspect « sémantique » dans la mesure où l'interprétation est privilégiée à la forme. Si le choix des critères est lié à la tâche, les critères sont en eux-mêmes génériques et indépendants du corpus et de la tâche. Une amélioration possible et intéressante consisterait en une démarche plus interactive avec l'apprentissage. En effet, le choix des critères est effectué manuellement et pour leur pertinence linguistique : une phase préliminaire qui étudierait et analyserait les n-grammes et / ou les segments discontinus révélateurs d'une classe pourrait faire émerger des critères supplémentaires.

Les performances de la méthode utilisant les séquences maximales de mots sont décevantes. Ainsi que nous l'avons évoqué à la section 5.2.2, l'apprentissage a été effectué dans de mauvaises conditions, ce qui a provoqué un biais important pour les performances attendues. Nous souhaitons retravailler sur cette méthode afin d'obtenir des résultats plus représentatifs.

Références

AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H. & VERKAMO A. (1996). Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, p. 307–328.

AHONEN-MYKA H. & DOUCET A. (2005). Data mining meets collocations discovery. In *Inquiries into Words, Constraints and Contexts, Festschrift for Kimmo Koskenniemi*, p. 194–203 : CSLI Studies in Computational Linguistics. .

BILHAUT F. & WIDLÖCHER A. (2006). LinguaStream : An Integrated Environment for Computational Linguistics Experimentation. In *11th Conference of the European Chapter of the Association of Computational Linguistics (EACL'06)*, p. 95–98.

CHAURAND J. & MAZIÈRE F. (1990). *La définition*. Larousse, collection Langue et Langage.

DOUCET A. (2004). Utilisation de séquences fréquentes maximales en recherche d'information. In *Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data (JADT-2004)*, p. 334–345, Louvain-La-Neuve, Belgium : JADT-2004.

DOUCET A. & AHONEN-MYKA H. (2006). Fast extraction of discontiguous sequences in text : a new approach based on maximal frequent sequences. In *Proceedings of IS-LTC 2006, Information Society - Language Technologies Conference*, p. 186–191, Ljubljana, Slovenia.

LI W., HAN J. & PEI J. (2001). Cmar : Accurate and efficient classification based on multiple class-association rules. In *IEEE International Conference on Data Mining (ICDM'01), San Jose, USA*.

RIOULT F., ZANUTTINI B. & CRÉMILLEUX B. (2008). Apport de la négation pour la classification supervisée à l'aide d'associations. In *Conférence francophone sur l'apprentissage automatique*.

STUBBS M. & BARTH I. (2003). Using recurrent phrases as text-type discriminators : A quantitative method and some findings. *Functions of Language*, **10**, 61–104(44).

VERNIER M., MATHET Y., RIOULT F., CHARNOIS T., FERRARI S. & LEGALLOIS D. (2007). Classification de textes d'opinions : une approche mixte n-grammes et sémantique. In *3ème DÉfi Fouille de Textes (DEFT'07) associé à la plateforme AFIA'07*, p. 99–109.

WIDLÖCHER A., BILHAUT F., HERNANDEZ N., RIOULT F., CHARNOIS T., FERRARI S. & ENJALBERT P. (2006). Une approche hybride de la segmentation thématique : collaboration du traitement automatique des langues et de la fouille de texte. In *Deuxième DÉfi de Fouille de Textes (DEFT'06), Semaine du Document Numérique (SDN'2006)*.