Une approche hybride de la segmentation thématique : collaboration du traitement automatique des langues et de la fouille de texte

Antoine Widlöcher, Frédérik Bilhaut, Nicolas Hernandez, François Rioult,
Thierry Charnois, Stéphane Ferrari et Patrice Enjalbert

GREYC, Université de Caen, CNRS UMR 6072
Campus Côte de Nacre, Bd Maréchal Juin, BP 5186 F-14032 Caen Cedex
{awidloch, fbilhaut, nhernand, frioult, charnois, ferrari, patrice}@info.unicaen.fr

Résumé. Dans cet article, nous présentons un travail mené au sein de l'équipe « Données Document Langue » du GREYC autour de la tâche de segmentation thématique de textes du DÉfi Fouille de Textes 2006. La méthode proposée combine des méthodes de traitement automatique des langues et de fouille de données : elle associe la détection de marqueurs discursifs génériques à une technique de fouille séquentielle de texte. Les différents traitements sont intégrés à l'aide de la plate-forme LinguaStream.

1 Introduction

Dans le domaine très général du traitement automatique des textes écrits, l'analyse thématique peut être définie comme s'attachant à la description du contenu informationnel des documents, ainsi qu'à sa distribution inter- et/ou intra-documentaire. Parmi les sous-problèmes soulevés par cette question, apparaît celui de la segmentation thématique, qui vise plus particulièrement à établir une certaine structuration des textes en termes de segments dits « thématiques », unités textuelles présentant une certaine homogénéité sur le plan thématique. Les critères caractérisant cette homogénéité sont plus ou moins bien définis selon les cas, mais un « consensus mou » semble s'être formé autour de la notion d'« à propos » : un segment thématique se constituerait autour d'une identité de « sujet », au sens de « ce dont on parle ». De fait, on conçoit aisément l'intérêt applicatif de cette approche sur le plan de l'accès assisté à l'information.

Cette tâche de segmentation thématique fait l'objet du deuxième DÉfi Fouille de Textes (DEFT'06)[1]. Il s'agit plus précisément de reconnaître automatiquement des segments thématiques dans des textes écrits en français et appartenant à différents domaines : politique, juridique et scientifique. Il nous est ainsi demandé de déterminer (directement ou indirectement) les premières phrases de chaque segment. La méthode est mise au point sur des corpus d'entraînement représentant 60% de chaque corpus et testée sur le reste (l'usage de corpus d'entraînement extérieur est interdit). Un segment consiste ici en une unité de texte constituée d'une ou plusieurs phrases successives. La notion de segment thématique est ici propre à chaque corpus : pour le corpus politique, unités de l'ordre du paragraphe tel que décidé lors de leur écriture ou lors de

[1] http://www.lri.fr/ia/fdt/DEFT06

la constitution des corpus mis en ligne par l'organisme en charge de cette tâche ; pour le juridique, une loi de l'Union Européenne ; dans l'ouvrage scientifique, chaque unité contiguë, une fois la structure logique retirée (des chapitres à la sous-sous-section).

Ce problème de segmentation thématique intra-documentaire est traditionnellement abordé par des méthodes dites distributionnelles ou statistiques, qui se fondent plus ou moins explicitement sur la notion de cohésion lexicale (Halliday et Hasan 1976). Il s'agit notamment de la lignée du *text-tiling* de Hearst (1997), qui a bien-sûr connu depuis de nombreux développements. Les méthodes de cette famille procèdent à une segmentation le plus souvent linéaire selon des critères quantitatifs. Elles s'inspirent de méthodes de recherche d'information en appliquant à l'intérieur du document des méthodes d'analyse de la distribution des mots habituellement appliquées au niveau inter-documentaire. Et en faisant comme elles l'économie de ressources spécifiques tout en mettant en œuvre des algorithmes relativement peu coûteux, ces méthodes bénéficient généralement d'une grande généricité et d'une applicabilité à grande échelle.

En revanche, la nature des segments thématiques qu'elles délimitent reste mal définie, dans la mesure où la notion de cohésion lexicale ne suffit pas à caractériser linguistiquement un hypothétique objet textuel que l'on appellerait « segment thématique ». Au contraire, il semble que cet objet soit défini *de facto* par les moyens mis en œuvre pour le délimiter automatiquement, plutôt que par un modèle posé *a priori*. Certes, la cohésion lexicale constitue vraisemblablement un indice très important sur le plan thématique, présentant de surcroît l'avantage d'être assez facilement accessible à l'analyse automatique. Mais il est tout aussi clair que la réalité de l'organisation thématique d'un texte fait intervenir bien d'autres critères, ce qui motive d'autres approches qui privilégient la modélisation linguistique des objets textuels étudiés.

Il s'agit alors d'exploiter des marqueurs ou indices porteurs d'indications de la structure thématique textuelle. Un exemple est donné par la notion de cadre de discours (Charolles 1997), structure dont la fonction thématique ne peut être ignorée tout en demeurant orthogonale au phénomène de cohésion lexicale. On peut par exemple s'intéresser aux cadres dits thématiques (Porhiel 2001), organisationnels (Jackiewicz et Minel 2003) ou spatio-temporels (Bilhaut et al. 2003), ou encore aux univers notionnels (Bilhaut et Enjalbert 2005). Sur un autre plan, les marqueurs de relations anaphoriques jouent également un rôle, (Smolczewska et Lallich-Boidin 2004) les exploitant parmi d'autres indices pour segmenter les documents techniques en unités thématiquement homogènes. À un niveau plus général, la prise en compte des titres s'avère également être un critère de structuration des plus pertinents (Laignelet 2006), de même que les connecteurs de discours tels qu'utilisés dans (Hernandez et Grau 2005) ou les phénomènes de parallélisme discursif (Guégan et Hernandez 2006). Plus généralement, même si elles visent à résoudre des sous-problèmes relativement spécifiques qu'il conviendrait d'intégrer au sein d'une approche plus globale, chacune de ces méthodes procède bien à la définition linguistique *a priori* de son objet d'étude, pour ensuite élaborer des systèmes d'analyse automatique. On peut d'ailleurs parfaitement considérer l'analyse du phénomène de cohésion lexicale sous cet angle, en la combinant à l'analyse d'autres objets discursifs, à la manière de (Ferret et al. 2001) au sujet des cadres thématiques.

Cependant, dans le cadre de la tâche fixée à l'occasion de ce défi de fouille de textes, l'approche linguistique « pure » est difficilement applicable car les « unités

thématiques » à délimiter sont de natures très différentes selon les corpus. Il semble impossible de les unifier au sein d'une seule et même modélisation linguistique, un article de loi n'ayant pas les mêmes propriétés qu'un segment de discours politique ou qu'un chapitre d'ouvrage scientifique[2].

En revanche, il est clair que les différents indices discursifs précédemment cités (introducteurs de cadres, titres, reprises anaphoriques, cohésion lexicale, etc.) sont bien présents et exploitables dans ces corpus comme dans tout autre texte. Il est seulement impossible de définir *dans un modèle unique* quel pourrait être leur rôle dans la délimitation des différents segments thématiques ici recherchés. Une approche par apprentissage s'impose naturellement dans ce contexte, puisque l'on peut ainsi espérer obtenir de façon automatique, et donc à moindre coût, un modèle spécifique à chaque corpus.

Se pose alors la question de la généricité des traits fournis au système d'apprentissage. Nous avons délibérément choisi, afin de conserver l'avantage d'une large applicabilité du système, *de n'utiliser que des marqueurs génériques*, présents dans l'ensemble du corpus fourni pour la tâche, et vraisemblablement pour une grande part des textes écrits en français[3]. La méthode se répartit ainsi en trois phases :

- reconnaissance automatique d'un certain nombre de marqueurs discursifs génériques, jugés pertinents sur le plan thématique ;
- apprentissage par une méthode de fouille de données séquentielle sur un corpus homogène ;
- application du modèle obtenu sur le corpus effectif.

Si le procédé est relativement classique, la pluridisciplinarité de notre équipe de recherche nous a permis de combiner des compétences en Traitement Automatique des Langues (TAL) et en Fouille de Données (FdD). Notre réflexion méthodologique s'est donc orientée vers la manière d'exploiter les caractéristiques discursives des textes (linéarité, marqueurs d'organisation de l'information, etc.) pour accomplir la tâche, et ce aussi bien pour le choix des descripteurs que pour le choix des techniques de fouille à utiliser. On peut de ce fait considérer l'approche proposée comme hybride, tout d'abord parce que les traits fournis au système de fouille résultent de procédés de TAL non triviaux, les indices que nous exploitons ne se résumant pas à des marques de surfaces exhaustivement énumérables. D'autre part, la valeur linguistique des modèles obtenus sera par la suite évaluée, et éventuellement intégrée pour partie aux modèles opérationnels que nous développons par ailleurs.

La description sommaire des procédés permettant de détecter dans les textes un certain nombre d'indices discursifs fera l'objet de notre première partie. La partie suivante sera consacrée aux méthodes de fouille spécifiquement adaptées à la nature textuelle des données, et permettant d'obtenir automatiquement un modèle des phrases de rupture thématique. La troisième partie décrira les différentes modalités de mise en œuvre automatique. Dans la dernière partie seront présentés et discutés les résultats

[2] Au sein même de ce dernier corpus, on peut même supposer que les différents segments à délimiter (chapitres, sections, textes intermédiaires, etc.) ne présenteront pas non plus les mêmes propriétés (et leur modélisation ferait en tout état de cause appel à des éléments de l'architecture textuelle, les titres en l'occurrence, qui ont été retirés du corpus).

[3] À l'exception, comme souvent, des genres littéraires où la question de l'analyse thématique est elle-même d'une autre nature.

obtenus par cette méthode hybride.

2 Méthodes de traitement automatique des langues

Avant de détailler les différents indices et les différentes marques retenus comme objets textuels ou comme propriétés d'objets textuels susceptibles d'indiquer la présence d'une rupture thématique, envisageons tout d'abord une formulation plus générale, en termes de « fonction de cohésion discursive », de la nature des fonctions textuelles recherchées, fonctions pouvant garantir leur pertinence pour une segmentation. La table 1 en fin de section récapitule les principaux indices linguistiques utilisés.

2.1 Différentes fonctions de cohésion discursive

Pour la détermination des indices à utiliser, nous nous appuyons sur une notion de « fonction de cohésion discursive » visant à subsumer, sous un ensemble restreint de catégories fonctionnelles générales, la diversité des opérations particulières qu'ils « déclenchent ». Sans nous prononcer ici sur la nature exacte des opérations effectivement mises en œuvre, aux niveaux textuel, sémantique, ou cognitif, nous pouvons cependant mettre en valeur certaines catégories assez fondamentales dont les indices retenus dans notre approche sont représentatifs. La majeure partie des indices présentés ci-après pourra être aisément rapportée à ces différentes catégories fondamentalement liées à notre tâche de segmentation.

Certains indices ont tout d'abord pour effet de marquer une *rupture*, ou *discontinuité*, dans le continuum textuel, faisant office de frontière entre des éléments rendus ainsi hétérogènes, non en vertu de leurs qualités intrinsèques, mais à la lumière du type de connexion introduit entre eux par la marque considérée.

D'autres marques ont par ailleurs pour fonction d'*introduire* ou d'*amorcer* une zone textuelle, c'est-à-dire de délimiter sa borne gauche, par exemple par une annonce thématique ou l'introduction d'un critère d'interprétation valable jusqu'à nouvel ordre.

Symétriquement, certains indices manifestent au contraire la *clôture* d'un segment textuel par l'indication de sa borne droite. À un niveau de granularité élevé, on pensera par exemple à la fermeture d'un développement par la concentration conclusive de son propos, venant clore la « phrase » discursive courante.

Enfin, d'autres objets textuels ont pour vocation de *prolonger* ou de *relayer* une information déjà introduite et dont on maintient ainsi la présence ou la saillance dans le texte. De telles marques de continuation peuvent plus précisément conduire à une réactivation d'un élément de sens, comme par exemple dans le cas de l'anaphore, soit, de façon plus diffuse, consister en un maintient d'homogénéité, par exemple lexicale.

2.2 Marqueurs discursifs

Connecteurs et expressions indicatives. Les connecteurs (« donc », « parce que », « ensuite », etc.) comme les expressions indicatives (« Le problème est », « Un autre aspect est », « De la même manière », etc.) sont des marqueurs essentiels à l'analyse de la cohérence relationnelle d'un texte (Degand et Sanders 2002, Knott 1996,

Schiffrin 1987). Inspiré de (Chartrand 1999) et en continuité des travaux de (Hernandez et Grau 2005), nous avons manuellement catégorisé 188 marqueurs[4] selon trois fonctions organisationnelles du discours. Les marques indiquant : *un commencement ou une amorce d'une unité de texte* précédant au moins une marque de continuation ou de terminaison ; *une continuation d'une unité* précédemment débutée (ou éventuellement déjà continuée) sans pour autant la terminer ni interdire sa continuation ; *la terminaison d'une unité ou d'une liste énumérative* interdisant toute continuation possible.

Lors de la phase de repérage des marques, celles apparaissant en tête de phrase ont été privilégiées. Cette contrainte repose sur l'hypothèse que cette position dans la phrase joue un rôle de premier plan dans l'organisation discursive.

Marqueurs anaphoriques. Parmi les éléments linguistiques susceptibles de marquer ou de signaler la structure du discours, l'anaphore, en tant qu'« expression référentielle non autonome », est souvent considérée comme la trace d'une continuité thématique (Piérard et al. 2004). Dans le cadre de la tâche qui nous occupe ici, ne semblent pertinentes que les anaphores dites extraphrastiques[5]. Nous les considérons comme telles, en première approximation, si elles sont en position préverbale dans la phrase. Parmi celles-ci, tous les types d'anaphore (pronominales, associatives, par reprise partielle de l'antécédent, etc.), sont considérés comme une marque potentielle de continuité thématique. Soulignons que le problème ardu de la résolution anaphorique n'est pas traité ici : la reconnaissance de la marque anaphorique (pronom personnel[6] de la troisième personne, pronom démonstratif, article défini, adjectif possessif et démonstratif) permet d'étiqueter toutes les marques potentielles d'anaphore.

Longueur des phrases. Elle est susceptible de fournir un premier indice significatif, si du moins l'on considère l'appartenance des trois corpus à des genres particuliers, expositifs et assez procéduraux, au sein desquels les différents segments sont souvent introduits par des tournures ou des motifs de tournure assez canoniques. Nous assouplissons ici la prise en compte de cet attribut en ne comptabilisant que les mots pleins effectivement présents.

Cohésion lexicale. L'exploitation du phénomène de cohésion lexicale pour la délimitation de zones textuelles est assez standard en matière de segmentation automatique, comme c'est le cas par exemple avec les approches de type *text-tiling*. Ici, la solution retenue propose une méthode simplifiée du calcul de cohésion lexicale, au sein de laquelle nous nous limitons au niveau phrastique pour mesurer le degré de cohésion d'une phrase avec la phrase précédente.

Introducteurs de cadre. S'appuyant sur les phénomènes d'encadrement du discours, une autre étape de l'analyse vise la détection d'introducteurs de cadres, expressions dont la fonction consiste à fixer un critère d'interprétation valable pour un segment textuel de taille variable, correspondant ce faisant à une possible marque de rupture thématique.

Temps verbaux. Leur analyse répond à plusieurs attentes. Dans la perspective restreinte des approches de type « encadrement », elle entre en jeu dans la résolution de la délicate question de la délimitation de segments pour lesquels un critère d'in-

[4]Issus de la traduction des ressources anglaises existantes et étendus à l'aide de dictionnaire commun de synonymes (Knott 1996, Marcu 1997).

[5]Anaphores qui réfèrent à un antécédent situé dans une phrase précédente.

[6]Nous opérons un filtrage des occurrences impersonnelles du pronom « il ».

terprétation temporel a été donné. Dans ce cas, le changement de temps verbal traduit fréquemment une fermeture de cadre et peut, à ce titre, être considéré comme une marque de rupture. D'une manière plus générale l'hétérogénéité du temps verbal constitue un indice important de discontinuité pouvant être pris en considération en tant que tel.

Phrases nominales. Nous procédons par ailleurs au marquage des phrases nominales, afin d'isoler les éléments textuels à valeur fortement structurante (titres, pseudo-titres, amorces d'énumération, etc.) et correspondant souvent à des indices de rupture.

Négation. La prise en considération de la négation répond à l'hypothèse selon laquelle une certaine densité de tournures négatives pourrait traduire la présence d'éléments problématiques ou critères de « problématisation », eux-mêmes souvent indices d'ouverture ou de fermeture, et donc liés aux mécanismes de rupture textuelle.

Ponctuation. Nous considérons également comme pertinent le relevé des ponctuations, et en particulier des tournures interrogatives qui traduisent potentiellement, elles-aussi, soit une problématisation, soit du moins la nécessité d'ouvrir un nouvel espace de réflexion. Une fois de plus, c'est la récurrence des marques de ce type qui pourra seule, à l'occasion, révéler la présence de discontinuités.

Critères énonciatifs. Enfin, d'un point de vue plus méta-textuel et pragmatique, l'incursion du locuteur/orateur et du lecteur/auditeur dans le texte, et plus particulièrement leur manifestation à travers l'utilisation de deux premières personnes du singulier et du pluriel coïncident fréquemment, au sein de textes aussi peu « dialogués » que ceux du corpus, à des « moments » du discours où la structure globale doit être mise en évidence. Ces éléments, tout à fait significatifs d'une point de vue rhétorique et argumentatif, peuvent à bon droit être considérés comme des éléments visant à expliciter et à faire apparaître différents points de rupture dans le flot textuel.

3 Méthode de fouille de données

Lorsque l'on considère chaque phrase de corpus comme un n-uplet, la caractérisation des débuts de segments thématiques est un problème de classification supervisée. En effet, nous disposons d'un corpus d'entraînement où chaque n-uplet est étiqueté, et d'un corpus de test dans lequel il s'agit de reconstituer la valeur de la classe pour chaque objet.

Pour résoudre ce type de problème, notre expertise concerne l'extraction de motifs ou de règles et leurs usages en classification supervisée. Le processus de fouille de données, appliqué aux corpus étiquetés par les méthodes de traitement de la langue, est décomposé en trois phases : la première concerne la construction de matrices de données booléennes, nécessaires pour la découverte de motifs. Ensuite des règles séquentielles sont extraites. Une méthode de vote clôt ce dispositif en déclenchant ces règles sur les corpus de test et fournit la décision finale.

3.1 Prétraitement des données

Les méthodes à base de motifs que nous utilisons requièrent des données booléennes : les attributs quantitatifs doivent être discrétisés. C'est le cas des attributs de cohésion

nom	type
classe	début, suivant début, intérieur, précédant fin ou fin de segment
cohésion lexicale	$[0, 1]$
nombre de mots pleins	numérique
phrase nominale	booléen
connecteur de début	0, 1, 2, 3+
connecteur de continuation	0, 1, 2+
connecteur de terminaison	0, 1, 2+
temps présent	0-1, 2, 3, 4-5, 6+
temps passé	0, 1, 2+
temps futur	0, 1, 2+
marque anaphorique	0, 1, 2, 3+
marque anaphorique en début de phrase	booléen
introducteur de cadre	booléen
forme impersonnelle	booléen
négation	0, 1, 2+
interrogation	booléen
exclamation	booléen
2ème personne	booléen
1ère personne	booléen

TAB. 1 – Liste des indices linguistiques utilisés.

lexicale et du nombre de mots. Leur distribution a été analysée sur l'ensemble du corpus d'apprentissage étiqueté. La cohésion lexicale a été discrétisée en six intervalles de population homogène :

$$[0, 0],]0, 0.09],]0.09, 0.15],]0.15, 0.23],]0.37, 1[, [1, 1]$$

Le nombre de mots a été divisé en trois intervalles : $[1, 6],]6, 14],]14, 196]$. Tous les autres attributs sont catégoriques. On dispose également de cinq valeurs de classe possibles pour une ligne : début de segment, suivante du début, intérieur du segment, précédente de la fin, ligne de fin.

À l'issue de la préparation des données, les matrices booléennes comportent 60 colonnes.

3.2 Extraction de règles séquentielles

Une méthode de classification supervisée à base de règles d'association, comme CMAR (Li et al. 2001), est inopérante car les classes sont très déséquilibrées. Il y a en effet une grande majorité de lignes intérieures à un segment, qui génèrent une grande quantité de règles concluant sur cette classe. Il est difficile de paramétrer un système de vote dans ces conditions. De plus, ce type de méthode ne prend pas en compte la séquentialité du discours.

La méthode proposée utilise alors des règles séquentielles. Elles sont constituées d'une prémisse représentant une succession d'attributs et d'une conclusion qui est un attribut. Traditionnellement, les règles sont plutôt construites à partir de séquences fréquentes de *motifs* (Masseglia et al. 2004). Considérant des attributs notés de A à C, un exemple de séquence est $< (AB)(A)(BC) >$. Elle signifie que l'on trouve une quantité significative de successions de trois objets, le premier contenant le motif AB, le deuxième A et le dernier BC. On en déduira une règle $< (AB)(A) > \rightarrow (BC)$.

Cependant, les classiques extracteurs de séquences sont inadaptés pour le problème courant, car ils ne contraignent pas l'écart entre deux motifs d'une séquence. Or la caractérisation précise du début de segment thématique nécessite des règles déterminant la présence de la conclusion dans l'objet suivant *immédiatement* le dernier élément de la prémisse. Pour satisfaire cette contrainte, nous avons utilisé l'extracteur de règles séquentielles WinMiner (Méger et Rigotti 2004) qui permet de limiter l'écart temporel entre les éléments d'une séquence.

Les règles fournies par WinMiner utilisent des succession de simples attributs et non pas de motifs d'attributs comme évoqué ci-dessus. Nous pensons malgré tout que cette restriction n'est pas rédhibitoire. Si la séquence $< (AB)(A)(C) >$ est fréquente, WinMiner calculera indépendamment les règles $< (A)(A) > \rightarrow C$ et $< (B)(A) > \rightarrow C$. La différence entre les enchaînements de motifs ou d'attributs réside dans l'élaboration du classifieur faisant voter les règles (*cf.* section 3.3).

Pour éviter une prépondérance de votes sur la classe « phrase interne », les règles concluant sur cette classe ne sont pas recherchées. De plus, tous les attributs exprimant l'absence d'une propriété sont ignorés[7]. Ces attributs sont en effet trop majoritaires et ne permettent pas une discrimination positive des lignes de rupture.

Nous utilisons des règles issues d'un type de corpus d'entraînement, parmi juridique, politique ou scientifique, pour classer les lignes du corpus de test de même type. Le support minimal est de dix objets pour un extrait de 3000 lignes du corpus juridique, de 2200 lignes en politique, et du corpus scientifique complet. La confiance minimale est de 15 %. Parmi les dix premiers extraits des corpus d'entraînement politique et juridique, nous retenons les règles de l'extrait ayant les meilleures performances sur notre plate-forme de test.

3.3 Méthode de décision

L'extracteur WinMiner fournit des règles séquentielles caractéristiques des corpus étudiés. Pour déterminer le type d'une phrase, nous avons réalisé un classifieur à l'aide d'une méthode de vote sur les règles extraites.

La phase de décision examine la succession des phrases du corpus de test. Lorsqu'une règle contient dans sa prémisse des attributs qui correspondent à cette succession, la conclusion est votée selon la confiance de la règle. Pour résoudre les inévitables conflits, les règles déclenchées sont regroupées par conclusion commune et la plus grande des moyennes des confiances détermine la classe d'une phrase.

[7]Ce procédé a été mis en œuvre indépendamment de la nature de l'attribut, ce qui réduit considérablement l'intérêt du trait « reprise anaphorique », dont c'est principalement l'absence qui est ici significative. Cette limitation sera corrigée à court terme.

Le discours induit une contrainte sur les séquences de classes successives. Par exemple, une prédiction « suivante de début » doit immédiatement suivre une prédiction « début ». La décision finale pour une phrase de rupture n'est donc rendue que si le résultat obtenu est cohérent avec celui de la phrase précédente.

Les règles concluant sur une valeur de classe sont les plus naturelles à mettre en œuvre et donnent les meilleurs résultats. Elles ont été utilisées pour fournir notre première proposition de résultats[8]. Pour la deuxième proposition, nous avons tenté d'utiliser les règles comportant une valeur de classe dans la prémisse. Nous avions l'intuition qu'il pouvait être profitable d'utiliser non seulement l'information précédant une phrase à classer mais également celle des phrases lui succédant. Cependant, cette méthode est peu efficace : les règles fournies par WinMiner ont en effet un maximum de confiance dans un intervalle donné. Cet extracteur n'est donc pas conçu pour produire des règles apportant de l'information sur un objet intermédiaire de la séquence. Notre troisième proposition de résultats consiste en l'intersection des deux premières.

4 Architecture et mise en œuvre

L'architecture que nous proposons pour la mise en œuvre des traitements indiqués s'appuie sur la plate-forme LinguaStream[9] (Widlöcher et Bilhaut 2005, Bilhaut et Widlöcher 2006) et tire avantage d'un certain nombre de ses principes.

Plate-forme générique dédiée au TAL, LinguaStream a été initialement développée pour faciliter la réalisation d'expériences sur corpus, ainsi que le cycle d'évaluation/ajustement qui en découle. Elle permet de mettre en œuvre des procédés non triviaux en facilitant la conception et l'évaluation de chaînes de traitements complexes, par assemblage visuel de modules d'analyse de types et de niveaux variés : morphologique, syntaxique, sémantique, discursif, etc. Chaque palier de la chaîne de traitement se traduit par la découverte et le marquage de nouvelles informations, sur lesquelles pourront s'appuyer les analyseurs subséquents. Un environnement de développement intégré permet de construire ces chaînes de traitement, à partir d'une « palette » extensible de composants. Certains sont spécifiquement dédiés à des traitements d'ordre linguistique, et d'autres permettent de résoudre différents problèmes liés à la gestion des documents électroniques, à la réalisation de calcul sur les annotations produites, etc. D'autres encore permettent de visualiser les documents analysés et leurs annotations.

Cependant, si l'objectif clair de la plate-forme est de proposer un « atelier » pour le TAL, et de fournir en standard un grand nombre de composants à vocation linguistique, les éléments du système dédiés à la mise en œuvre de chaînes de traitement sont génériques, et il est donc tout à fait possible d'intégrer à une chaîne des traitements de natures variées. Ainsi, un ensemble de composants dédiés à la fouille de données ont pu être facilement mis en œuvre, souvent par simple établissement d'un pont vers une application externe.

[8] Les résultats présentés plus loin sont ceux obtenus par cette méthode.
[9] `http://www.linguastream.org`

4.1 Mise en œuvre de l'analyse linguistique

Chaque composant ou ensemble de composants vise ainsi l'analyse et le marquage d'un type d'objet linguistique donné. Pour procéder à cette analyse sans recourir à des implémentations *ad hoc* au sein desquelles la compétence linguistique se trouverait « dispersée », nous utilisons, dès que possible, pour garantir la capitalisation de cette expertise, les formalismes déclaratifs proposés par la plate-forme pour l'écriture de règles à différents niveaux de granularité et selon différentes approches. Nous nous appuyons ici en particulier sur l'écriture de grammaires locales d'unification EDCG (*Extended Definite Clause Grammar*), de patrons par transducteurs MRE (*Macro Regular Expressions*) et de grammaires de contraintes CDML (*Constraint-based Discourse Modeling Language*).

En fin de chaîne, nous disposons ainsi d'un corpus au sein duquel sont délimités les différents objets textuels retenus pour notre approche, auxquels sont associées différentes annotations correspondant aux attributs que nous souhaitons faire ressortir. Un dernier composant permet de produire une matrice au sein de laquelle chaque phrase est représentée par une ligne, et chaque colonne par un attribut pouvant correspondre soit à un type d'objet textuel annoté, soit à une propriété associée à ce type d'objet, soit à une propriété associée à la phrase elle-même. Les valeurs de chaque attribut correspondent, selon les cas, à un nombre d'occurrences de l'objet ou de la propriété, ou à une valeur calculée pour la propriété, comme par exemple pour la cohésion lexicale. La discrétisation et/ou la binarisation des valeurs sera réalisée au sein de la chaîne de traitement dédiée à la fouille de données (*cf.* section 3.1).

La figure 1 représente la chaîne LinguaStream utilisée pour les traitements d'ordre linguistique.

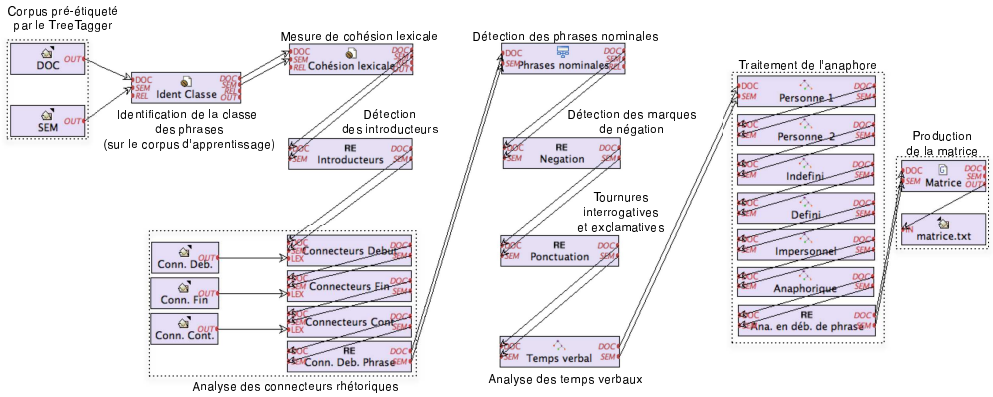

FIG. 1 – La chaîne de traitement linguistique.

4.2 Mise en œuvre de la méthode de classification

L'ensemble du travail décrit aux sections 3.1 à 3.3 a également été réalisé avec la plate-forme. Elle permet d'éviter l'assemblage traditionnel de scripts et fournit une fois encore une représentation graphique de l'enchaînement des tâches. Cette représentation est très explicite, tant à des fins de communication que de mémoire du traitement et de ses paramètres.

La figure 2 présente la chaîne réalisant l'extraction de règles séquentielles, décrite à la section 3.2. Chaque ligne extrait les règles relativement à chaque classe. Le premier traitement (effectué par l'utilitaire Unix `sed`) ne conserve que la classe souhaitée. Le second traitement (`bin2seq`) transforme les formats binaires en séquentiel. L'étape WinMiner extrait les règles. Seules les règles concluant sur une valeur de classe sont conservées (`mkrulseq`) ou qui contiennent une valeur de classe dans la prémisse (`mkseq`). Le nom des fichiers d'entrée et de sortie est paramétrable, à l'aide de la variable `NOM` suivie d'une extension particulière.

L'interface de LinguaStream sert ici à la mise au point des traitements. Pour la production des résultats complets sur le corpus de test, ces traitements ont été appliqués en mode *runtime*, une commande système qui applique la chaîne avec les paramètres souhaités.

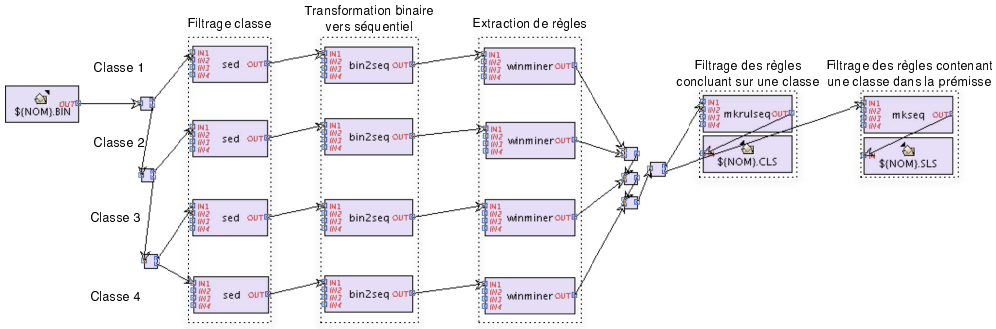

FIG. 2 – Extraction de règles.

5 Résultats et discussion

Si le caractère hybride de notre approche intégrant, en un même mouvement, analyse linguistique profonde et fouille de données nous semble prometteuse, si le caractère générique et léger, en termes de ressources, de la solution de TAL proposée nous semble pertinent, et si nous pouvons nous réjouir d'avoir mis en œuvre une architecture unifiée pour articuler ces différents traitements, reste que nos résultats (cf. tableau 2) pourraient être largement améliorés. Les différentes perspectives suivantes mériteraient à ce titre d'être considérées.

La qualité des résultats obtenus pourrait tout d'abord être améliorée en considérant

	Précision	Rappel	F-Score	Moyenne DEFT'06
Politique	24%	90%	38%	39%
Juridique	9%	25%	14%	26%
Scientifique	19%	91%	31%	29%

TAB. 2 – Résultats obtenus sur les différents corpus (F-Score souple, fenêtre 2).

un réservoir de règles plus représentatif du corpus. Pour des raisons d'efficacité algorithmique, la méthode présentée utilise en effet les règles obtenues sur une fraction de quelques milliers de lignes. Les règles découvertes sont donc certainement d'une portée restreinte pour un corpus si vaste et hétérogène.

La nature des règles utilisées est finalement assez restrictive : ce sont les règles séquentielles concluant sur une valeur de classe. Nos perspectives sur ce point concernent le calcul de règles permettant de classer une ligne selon les informations qui la précèdent mais également en mettant à profit les informations obtenues sur les lignes suivantes. Cette approche nécessite la mise au point d'un extracteur dédié.

Une autre perspective concerne l'exploitation de relations sémantiques qui peuvent jouer un rôle dans la structuration du discours sans être directement accessibles par des marques de surface. On peut par exemple mentionner le cas d'introducteurs de cadres qui ne peuvent être perçus comme tels que si on considère leurs interactions sémantiques avec d'autres référents du discours environnant (Bilhaut 2006). Le traitement de ces cas nécessite généralement de disposer de ressources dites « ontologiques », ce qui demeure problématique et entre éventuellement en conflit avec l'optique généraliste que nous nous sommes fixée. Cependant, il est parfaitement envisageable d'adopter une approche purement endogène pour constituer ces ressources. On peut notamment procéder à l'apprentissage sur corpus d'expressions dont la fonction discursive n'est significative que pour un domaine spécifique, et améliorer ainsi la segmentation thématique des textes du même domaine (Bilhaut et Enjalbert 2005).

D'un point de vue général, notre approche est d'emblée plutôt « sémantique » au sens où elle privilégie l'interprétation à la forme, en accordant un sens très fort à la dimension thématique, au dépens de critères plus surfaciques, mais peut-être efficaces, pouvant être obtenus, par exemple, par apprentissage. Ainsi, par exemple, le critère de non-rupture de chaîne anaphorique nous semble décisif, moins pour son respect systématique en corpus, qu'en vertu de sa valeur interprétative. L'inconvénient manifeste de cette approche, que nous assumons cependant, est qu'elle suppose la prise en compte de mécanismes interprétatifs complexes quand une solution de type apprentissage pourrait laisser émerger ces critères d'eux-mêmes.

Par ailleurs, nous nous appuyons ici sur un ensemble de descripteurs définis *a priori*, en vertu de leur pertinence linguistique. Cependant une étape préliminaire d'apprentissage sur un corpus auxquels auraient été appliqués des traitements de natures variées, sans *a priori* sur leurs rapports possibles avec la segmentation thématique pourrait faire apparaître la pertinence de descripteurs moins attendus. Une étude plus poussée du corpus après annotation des descripteurs retenus permettrait du reste une meilleure compréhension des mécanismes en jeu et une amélioration notable de l'ensemble retenu,

puis des règles qui en découlent.

Insistons également sur le fait que nous avons écarté pour l'heure les éléments ne pouvant faire l'objet d'une certaine « justification linguistique ». Ainsi, par exemple, une prise en compte de la longueur moyenne de segments pour un corpus donné permettrait peut-être d'améliorer les résultats en post-traitant l'application du classifieur. D'une manière générale, des points de vue plus statistiques et/ou distributionnels sur les données permettraient sans doute d'améliorer les résultats, mais de manière peut-être plus « aveugle ».

Dans (Hernandez et Grau 2005), l'utilisation de marques type connecteur selon une catégorisation *a priori* en classes organisationnelles apportait des résultats très pertinents. Néanmoins, cette approche ne résout pas le problème de l'ambiguïté d'une marque qui peut parfois appartenir à plusieurs classes. Nous envisageons d'adopter une approche qui consistera non plus à annoter les marques mais les énoncés d'un texte selon les différentes classes d'organisation. L'utilisation de techniques de fouille pourra ainsi faire apparaître les configurations de marques les plus caractéristiques d'une classe. Cette approche est bien sûr plus souple mais aussi plus coûteuse puisqu'elle requiert une activité d'annotation manuelle de corpus.

Pour l'heure, l'interface entre les traitements d'ordre linguistique et l'extraction des règles qui leur fait suite n'a par ailleurs pas encore donné lieu à de réelles interactions. Il conviendrait en effet à présent d'opérer un retour critique sur les règles extraites, afin d'en évaluer la qualité et d'améliorer le choix et la détection des traits utilisés. Si cette étape permet effectivement d'améliorer les résultats obtenus, nous espérons finalement intégrer certaines des règles obtenues au sein de modèles opérationnels que nous développons par ailleurs.

Enfin, rappelons que les traitements auxquels nous procédons sont rigoureusement les mêmes sur chacun des corpus proposés. Une spécialisation sur chacun d'entre eux, et donc l'injection de certains paramètres relatifs aux genres ou aux domaines considérés pourraient certes s'avérer efficace. Mais lorsque l'utilisation d'un système d'apprentissage est envisageable, il semble pertinent de conserver le bénéfice qui en découle en termes de généricité.

Références

Bilhaut, F., Ho-Dac, L.-M., Borillo, A., Charnois, T., Enjalbert, P., Le Draoulec, A., Mathet, Y., Miguet, H., Péry-Woodley, M.-P. et Sarda, L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. In Actes de la 10e Conférence Traitement Automatique du Langage Naturel (TALN'03), Batz-sur-Mer, France, pp. 315-320.

Bilhaut, F. et Enjalbert, P. (2005). Discourse Thematic Organisation Reveals Domain Knowledge Structure. In Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI'05), Pune, India, pp. 2815-2831.

Bilhaut, F. et Widlöcher, A. (2006). LinguaStream : An Integrated Environment for Computational Linguistics Experimentation. In Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (Companion Volume), Trento, Italy, pp. 95-98.

Bilhaut, F. (2006). Introducteurs d'univers intra-prédicatifs et leur détection automatique. In Actes du Colloque International Discours et Document, Caen, France, pp. 41-50.

Charolles, M. (1997). L'encadrement du dicours – Univers, champs, domaines et espaces. Cahiers de recherche linguistique 6.

Chartrand S-G., Aubin, D., Blain, R., Simard, C. (1999). Grammaire pédagogique du français d'aujourd'hui, Boucherville, Les Publications Graficor, pp. 397.

Degand L., Sanders T. (2002). The impact of relational markers on expository text comprehenion in L1 and L2, Reading and Writing, 15 (7-8), pp 739-757.

Ferret, O., Grau, B., Minel, J., Porhiel, S. (2001). Repérage de structures thématiques dans des textes, TALN, Tours.

Guégan M., Hernandez, N. (2006). Retrieving Textual Parallelisms. In Proceedings of EACL, Trento, Italie.

Halliday, M. A. K. et Hasan, R. (1976). Cohesion in English. London : Longman.

Hearst, M. A. (1997), TextTiling : Segmenting Text into Multi-Paragraph Subtopic Passages Computational Linguistics, vol. 23, pp. 33-64.

Hernandez, N., Grau, B. (2005). Détection automatique de Structures Fines du Discours, TALN, 6-10 juin, Dourdan.

Jackiewicz A. et Minel, J.-L. (2003). L'identification des structures discursives engendrées par les cadres organisationnels. In Actes de TALN'03 Batz-sur-Mer, France.

Knott, A. (1996). A Data-Driven Methodology for Motivating a Set of Coherence Relations Department of Artificial Intelligence, University of Edinburgh.

Li W., Han J. et Pei, J (2001). CMAR : Accurate and Efficient Classification Based on Multiple Class-Association Rules, IEEE International Conference on Data Mining (ICDM'01), San Jose, USA.

Laignelet, M. (2006). Les titres et les introducteurs de cadres comme indices pour le repérage de segments d'information évolutive. In Actes de ISDD 2006, Caen, France.

Masseglia F., Teisseire, M. et Poncelet, P. (2004). Extraction de motifs séquentiels, problèmes et méthodes, Extraction de motifs dans les bases de données, RSTI série ISI vol. 9, no 3-4.

Méger M., et Rigotti, C. (2004). Constraint-based mining of episode rules and optimal window size, In Proceedings of the ECML/PKDD'04 International Conference.

Piérard, S., Degand, L., and Bestgen Y. (2004). Vers une recherche automatique des marqueurs de la segmentation du discours. In Purnelle, G., Fairon, C., et Dister, A. éds. : Actes des 7e Journées internationales d'Analyse statistique des Données Textuelles, Louvain-la-Neuve, pp. 859-864.

Porhiel, S. (2001). Linguistic Expressions as a Tool to Extract Thematic Information. In Corpus Linguistics, pp. 477-482.

Schiffrin, D. (1987). Discourse Markers, Cambridge University Press.

Smolczewska, A. et Lallich-Boidin, G. (2004). Validation par prototypage d'un modèle de segmentation des documents techniques composites. In Enjalbert P. et Gaio, M., éds : Approches sémantiques du document numérique, Actes du septième Colloque International sur le Document Electronique (CIDE.7), La Rochelle, France, pp. 75-92.

Widlöcher, A. et Bilhaut, F. (2005). La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus. In Actes de TALN 2005, Dourdan, France, pp. 517-522.

Remerciements

Les auteurs remercient Nicolas Méger et Christophe Rigotti pour la mise à disposition de l'extracteur de règles séquentielles WinMiner. Cette recherche a bénéficié d'un soutien de l'État et de la Région Basse-Normandie dans le cadre d'un stage postdoctoral.

Summary

In this paper, we present a collaborative work led by the DoDoLa team from the GREYC Laboratory on the text segmentation task of the DEFT'06 challenge. The proposed method combines natural language processing methods and data mining techniques : we associate the detection of discursive cues with a sequential text mining method. Integration is achieved using the LinguaStream platform.