SATIM : Système d'Analyse et de Traitement de l'Information Multidimensionnelle

Ismaïl Biskri (*) (**), Jean-Guy Meunier (**)

(*) Université du Québec à Trois Rivières
Département de Mathématiques et d'Informatique
C.P. 500, Trois-Rivières (Québec), Canada, G9A 5H7

(**) Université du Québec à Montréal
Laboratoire d'ANalyse Cognitive de l'Information
C.P. 8888, succursale Centre-ville, Montréal (Québec), Canada, H3C 3P8

ismail_biskri@uqtr.ca
meunier.jean-guy@uqam.ca

Abstract

Diversity of the sources of information and the types of encoding of this information, and the variety of the sequences necessary to their processing with using of the various mathematical models that we explored in our preceding works required the creation of a data-processing platform : SATIM. This platform is adaptable, flexible, modular and allows the fast creation of a multitude of data processing sequences. It can also be increased by new modules. We will give in this article the details of platform SATIM. We will illustrate that with an example of a possible data processing sequence (in fact GRAMEXCO/eGRAMEXCO) conceived directly using SATIM.

Résumé

La diversité des sources d'informations et des types d'encodage de ces informations, ainsi que la variété des chaînes nécessaires à leur traitement et à l'application des différents modèles mathématiques que nous avons explorés dans nos précédents travaux ont nécessité la création d'une plate-forme informatique : SATIM. Cette plate-forme est adaptable, flexible, modulaire et permet la création rapide d'une multitude de chaînes de traitement. Elle peut également être augmentée par de nouveaux modules. Nous donnerons dans cet article les détails de la plate-forme SATIM. Nous illustrerons cela par un exemple d'une chaîne de traitement possible (en l'occurrence GRAMEXCO/eGRAMEXCO) conçue directement à l'aide de SATIM.

Mots clés : analyse multidimensionnelle, chaîne de traitement, unité d'information, multimédia.

1. Introduction

Depuis une décade, à peu près, le traitement de l'information a pris une ampleur qu'on ne soupçonnait certainement pas. D'abord textuel en raison des masses de documents textuels dont il fallait rendre compte du contenu dans des temps d'analyse raisonnables, le traitement de l'information est devenu un « traitement multimédia » avec l'essor de l'Internet. Seule une analyse textuelle à la fois rigide dans sa conception et limitée au seul traitement des textes écrits ne suffit plus. Un outil de veille concurrentielle, par exemple, mis entre les mains d'un utilisateur travaillant dans le domaine de la finance ou de la bourse ne devrait il pas permettre

non seulement l'analyse de données textuelles dans leur variété de codage (ascii, xml, html, etc.) et de langue mais aussi de données autres que du texte (son, image, vidéo, graphique, etc.)?

La littérature scientifique inspirée du TALN présente une multitude de travaux, de méthodes et d'approches qui proposent des traitements spécifiques sur des corpus textuels que nous pouvons classer, pour l'essentiel du moins, en deux grands groupes : logique et linguistique pour une première classe, numérique et statistique pour une deuxième classe. En ce qui concerne les outils numériques, leur apparente simplicité de conception dans la mise en œuvre de différentes chaînes de traitement (catégorisation, terminologie, classification, routing d'information, etc.) en fait une approche très cotée auprès des chercheurs qui s'intéressent à d'autres domaines que le TALN, en particulier l'informatique médicale, la reconnaissance de forme, etc. En effet, les approches numériques recherchent dans des corpus textuels les parties dans le texte qui sont semblables et les regroupent dans des classes de similarités. Ces deux opérations sont fondées sur le principe de la régularité de cooccurrence de chaînes de caractères qui sont des unités d'information statistiquement comparables et dont il est aisé de calculer les fréquences d'apparition dans les différents segments textuels. Ceci étant la même approche est possible pour d'autres types d'information (son, image, etc.). Reste bien sur à déterminer et à choisir la nature des unités d'information pour chaque source en entrée ainsi que le type de segmentation pour définir les parties à comparer. La définition d'une unité d'information (dans un sens non limité au seul contexte de la langue) est tributaire d'au moins quatre contraintes :

(i) Les unités d'information doivent être des portions du document en input ;

(ii) Il doit être facile sur le plan informatique de repérer les unités d'information ;

(iii) Les unités d'information doivent être statistiquement comparables. Il doit être aisé d'en calculer les fréquences d'apparition dans les différentes parties du document et par conséquent d'estimer leur distribution et la régularité à laquelle plusieurs unités cooccurrent dans les mêmes parties du document.

(iv) À l'instar de Balpe et al (1996), nous pensons que la définition d'une unité d'information dépend de l'objectif de lecture et de compréhension que nous nous donnons ainsi que l'usage qui en est attendu. Il en est de même pour le choix du type de segmentation qui est fortement lié au but poursuivi à travers la mise en œuvre de la chaîne de traitement.

2. Une plate-forme pour le traitement de l'analyse multidimensionnelle :

La diversité des sources d'informations et des types d'encodage de ces informations, ainsi que la variété des chaînes nécessaires à leur traitement et à l'application des différents modèles mathématiques que nous avons soit explorés soit développés, ont nécessité la création d'une plate forme informatique : SATIM (Système d'Analyse et de Traitement de l'Information Multidimentionnelle). Cette plate-forme est :

(i) Adaptable à toute sorte d'inputs, qu'ils soient textuels ou pas. La seule condition étant que l'input ait une configuration qui permette de dégager des segments et des unités d'information. Ces derniers serviront à établir la similarité entre les différents segments.

(ii) Modulaire. Chaque module réalise une tâche bien précise. Il peut être remplacé par un autre jugé plus performant dans la réalisation de la tâche. Par exemple dans le cas d'un texte l'extraction des unités d'information peut être opérée par un module qui isole les

mots (auquel cas l'unité d'information est le mot) ou encore par un module qui isole des chaînes de quatre caractères qui se suivent (auquel cas l'unité d'information est une chaîne de quatre caractères). L'usager final ainsi pourra choisir l'un ou l'autre des deux modules pour accomplir la tâche dont il est question. Les différents modules sont indépendants de la conception de la plate-forme. L'ajout d'autres modules peut augmenter la capacité de tâches à accomplir à l'aide de la plate-forme. Ce dernier point est particulièrement précieux en ce sens qu'il permet d'une part de remplacer des modules jugés peu performants (d'un point de vue technologique) et qu'il permet d'autre part de compléter des chaînes de traitement par d'autres modules que ce soit en amont ou en aval, et ce en vue de traitements encore plus complet, plus rigoureux, avec des résultats plus fins. Nous pensons pour ce dernier point, l'ajout de modules linguistique comme par exemple dans (Biskri, Delisle, 1999).

(iii) Flexible. La modularité poussée à son extrême fait qu'un minimum de rigidité affecte la plate-forme. Par un simple jeu de paramètres il sera possible à un utilisateur de créer assez rapidement une multitude de chaînes de traitement pour l'analyse numérique de divers types de données quelque soit leur codage ou leur source.

3. Une chaîne de traitement pour l'aide à l'extraction des connaissances dans des bases de données textuelles multilingues :

Le multilinguisme est un aspect du TAL qui ne cesse de se développer depuis l'avènement de l'Internet. Le challenge auquel nous faisons face est de savoir quel outil peut rendre compte de plusieurs langues sans pour autant subir de modifications majeures dans sa conception de sorte a l'adapter à chacune des langues. L'analyse multidimensionnelle dont c'est le propos dans cet article repose sur deux premiers points fondamentaux : la nature de l'unité d'information et la nature du segment. Nous avons eu l'occasion dans nos précédentes publications (Meunier, Biskri & al., 1997) (Biskri & Meunier, 1998) de présenter des modèles d'analyse numérique où l'unité d'information est le mot et le segment est toute portion du texte de n mots (le n étant un paramètre à saisir). Le pari à ce moment était de montrer la pertinence de l'approche dans des domaines comme la terminologie, le routing d'information etc., et ce pour la langue française. Bien entendu, dès lors que l'objectif devient une analyse multilingue, le module dédié à l'extraction des unités d'information change. Il est tout simplement remplacé par un autre fondé sur les n-grams de caractères. Le module dédié à la segmentation peut être également remplacé pour d'autres raisons. On en donnera un aperçu plus loin.

3.1. Les n-grams de caractères :

Bien qu'ayant été proposée depuis longtemps et utilisée principalement en reconnaissance de la parole, la notion de **n-grams de caractères** prit davantage d'importance avec les travaux de Greffenstette (1995) sur l'identification de la langue, de Damashek (1995) sur le traitement de l'écrit. Ils prouvèrent que ce découpage, bien que différent d'un découpage en mots, ne faisait pas perdre d'information. Parmi les applications plus récentes des n-grams on retrouve des travaux sur : l'indexation (Mayfield & McNamee, 1998) ; l'hypertextualisation automatique multilingue avec les travaux de Halleb et Lelu (1998) qui, à travers une méthode de classification thématique de grandes collections de textes, indépendante du langage, construisent des interfaces de navigation hypertextuelle ; ou encore l'analyse exploratoire multidimensionnelle en vue d'une recherche d'information dans des corpus textuels (Lelu *et al.*, 1998).

On définira un n-gram de caractères par une suite de n caractères : bi-grams pour n=2, tri-grams pour n=3, quadri-grams pour n=4, etc. Il n'est plus question de chercher un délimiteur comme c'était le cas pour le mot. Un découpage en n-grams de caractères, quelque soit n, reste valable pour toutes les langues utilisant un alphabet et la concaténation comme opérateur de construction de texte. Le choix des n-grams apporte un autre avantage très important : il permet de contrôler la taille du lexique et de la maintenir à un seuil raisonnable. La taille du lexique était jusqu'à présent l'aspect le plus controversé et considéré comme une limite des techniques fondées sur la comparaison des chaînes de caractères. En effet, un découpage en mots fait que la taille du lexique est d'autant plus grande que le corpus est grand. Cette limite subsiste malgré certains aménagements tels le "nettoyage" des mots fonctionnels, la lemmatisation et la suppression des hapax. Un lexique obtenu suite à un découpage en n-grams de caractères ne peut dépasser la taille de l'alphabet à la puissance n. Le choix d'un découpage en quadri-grams pour une langue de 26 caractères donnerait une taille maximale de 26^4 entrées, soit un lexique de 456 976 quadri-grams possibles. Si on élimine les combinaisons qu'il est impossible de rencontrer (p.ex. AAAA, ABBB, BBBA, etc.), ce nombre diminue de façon considérable. D'ailleurs ce nombre est estimé par Lelu *et al*. (1998) à quelques 13 087 quadri-grams pour un texte de 173 000 caractères.

Dans une approche avec découpage en n-grams de caractères, contrairement aux approches avec découpage en mots, il n'est pas question d'utiliser la lemmatisation pour réduire le lexique. La lemmatisation (qui consiste à remplacer une forme fléchie par son lemme) est, d'une part, relativement lourde à mettre en œuvre sur le plan informatique mais en plus, impose un traitement spécifique à chaque langue. Qui plus est, plusieurs lemmatiseurs ne semblent pas être en mesure de ramener des termes comme informatisation, informatique, et informatiser à un même concept qu'est l'informatique. Or souvent dans les corpus, on utilise des expressions ayant quasiment le même contenu informationnel comme, par exemple, dans les segments suivants : "l'informatisation de l'école", "informatiser l'école" et "introduire l'informatique à l'école". Le découpage des trois segments en n-grams est suffisant pour classer les trois segments dans la même classe car, outre le mot école qui est redondant dans les trois expressions, les tri-grams inf, nfo, for, orm, rma, mat et ati, permettent par un calcul de similarité d'affirmer que c'est d'informatique dont il est question. Par ailleurs, les tri-grams susmentionnés apparaissent aussi dans le découpage des mots information, informationnel, etc., ce qui peut être considéré à juste titre comme du bruit, à moins bien sûr que l'on évoque une interprétation sémantique particulière de l'informatique comme étant une science de l'information. On comprend dès lors tout l'intérêt d'une plate-forme flexible qui nous permettrait de lemmatiser sans pour autant nous y obliger. Le cas présent en est la meilleure preuve.

3.2. GRAMEXCO (les n-GRAms dans l'Extraction des Connaissances):

GRAMEXCO est un outil logiciel que nous avons développé pour la classification numérique des gros corpus et l'extraction de connaissances sur le contenu des textes. La classification numérique s'effectue au moyen d'un classifieur numérique comme ceux explorés dans (Meunier & al., 1997) (Biskri & Delisle, 1999) (Benhadid & al., 1998) (Rialle & al., 1998). L'unité d'information considérée est le n-gram de caractères, la valeur de n étant paramétrable. L'objectif visé est de fournir la même chaîne de traitement, peu importe la langue du corpus, avec toutefois des aménagements dans la présentation des résultats pour en permettre une relative facilité de lecture comme nous le verrons plus loin. Le fonctionnement de GRAMEXCO n'est pas totalement automatique. Le choix de certains paramètres est fait par l'utilisateur en fonction de ses propres objectifs. Du choix de ces paramètres dépend

l'interprétation des résultats qui se fait par l'utilisateur en fonction de sa subjectivité. GRAMEXCO prend en entrée un texte brut (non indexé) sous format ASCII[1]. Il s'en suit trois grandes étapes où l'utilisateur peut paramètrer certains traitements.

1. La **première étape** consiste à construire la liste des n-grams de caractères contenus dans le texte ainsi qu'à partitionner le corpus en plusieurs segments. Les deux opérations se faisant simultanément, nous récupérons en sortie une matrice où seront répertoriés les fréquences d'apparition de chaque n-gram dans les différents segments. Le choix de la valeur du n (bi-gram, tri-gram, quadri-gram, etc.) dépend de l'utilisateur et de l'expertise qu'il veut mener. Outre la valeur du n, d'autres paramètres sont la possibilité d'effectuer la conversion des caractères non alphanumériques en caractère espace, ou encore la conversion des chiffres en caractère espace. Ces deux paramètres répondent aux besoins d'une analyse pour laquelle les chiffres, la ponctuation ou encore d'autres caractères spécifiques seraient importants pour la qualité des résultats. Dans un texte technique par exemple, il serait peut être intéressant de savoir si version1 est différente de version2 et, par conséquent, les chiffres pourraient avoir autant d'impact informatif que les caractères alphabétiques. Le dernier paramètre pour les n-gram est en rapport avec la conversion des caractères majuscules en minuscules, ou vice versa. Si aucune de ces conversions n'est choisie, alors GRAMEXCO distinguera les lettres minuscules des majuscules. L'autre aspect important de cette première étape est le paramétrage de la segmentation. Ainsi, nous pouvons partager le texte soit en des sections formées d'un nombre déterminé de phrases, de paragraphes ou de mots, ou tout simplement des sections séparées par un caractère spécial. Ce paramètre est toujours choisi par l'utilisateur. Le pseudo-lexique formé de n-grams subit au cours de cette première étape un nettoyage soit, l'élimination des "n-grams hapax" dont la fréquence est inférieure à un certain seuil ou supérieure à un autre seuil, l'élimination de n-grams spécifiques sélectionnés dans la liste (par exemple des n-grams contenant des espaces) ou encore, si on veut pousser les choses plus loin, l'élimination de certains n-grams considérés comme fonctionnels, particulièrement les suffixes. Les paramètres décrits dans cette première étape sont représentés au niveau de l'interface usager par des boutons. L'action de chaque bouton étant bien entendu l'action d'un module. Rajouter d'autres modules pour cette première étape se traduira tout simplement par l'ajout d'autres boutons. La suppression d'un module se traduira par la suppression d'un bouton. Les paramètres sont pensés indépendants les uns des autres. Ainsi, choisir de distinguer les minuscules des majuscules par exemple n'impliquera pas forcément de choisir la suppression des n-grams hapax.
2. Dans la **deuxième étape**, les segments représentés dans la matrice obtenue à l'étape précédente sont comparés entre eux au moyen d'un classifieur numérique. Ce classifieur peut être un réseau de neurones (par exemple ART) (Meunier & al. 1997) ou un algorithme génétique (Rialle & al., 1998) ou encore fondé sur les champs de Markov cachés (Remaki & al., 2000), ainsi que d'autres classifieurs dont on retrouve une bonne rétrospective dans (Turenne, 2000). Le choix du classifieur devient un paramètre dans la conception de la chaîne de traitement. Le paramètre, étant défini, déclenche l'exécution du module de classification associé. Les segments qui sont semblables, étant donnée une certaine fonction de similarité (dépendant du classifieur choisi), seront classés dans les mêmes groupes. En simplifiant, on peut dire que deux segments sont semblables s'ils sont constitués des mêmes n-grams avec des fréquences presque identiques. Pour l'évaluation qui va suivre nous avons choisi ART. Ce choix n'est pas dicté par des

1.1.1.1.1.1.1.1.1.

[1] Ici, le module qui lit les données en entrée est conçu pour lire de l'ascii. Il est bien sur possible de le remplacer par un autre module qui lit des données dans un autre format que l'ascii.

raisons de performances particulières car tel n'est pas notre objectif. Nous aurions tout aussi bien pu choisir un autre classifieur qui aurait certes donné des résultats différents. De telles variations apparaissent dans les résultats d'une étude expérimentale et comparative des méthodes statistiques et des champs de Markov pour l'analyse de textes par ordinateur présentés dans Benhadid *et al*. (1998).

3. La configuration du résultat de la classification numérique se présente par l'affichage des classes de segments et, pour chaque classe, l'affichage des segments qui la constituent d'une part, et du lexique qui la forme d'autre part. À cette **troisième étape** la notion de n-gram n'est plus de mise. Il serait en effet impossible à un utilisateur d'interpréter des résultats et de donner des thèmes aux différentes classes à partir d'une seule liste de n-grams. Comme le souligne Turenne (2000), l'interprétation de telles classes est déjà un exercice non trivial en lui-même, dépendant *des* points de vue de l'utilisateur : il ne serait donc pas utile de lui rendre cette phase moins intuitive en utilisant une liste de n-grams. Le lexique de chaque classe est formé par les mots que contiennent les différents segments de cette classe. L'utilisateur pourra considérer le lexique comme l'union des mots des segments pour déterminer le thème global des classes, leur intersection pour déterminer le thème commun partagé par les segments, leur différence pour identifier des gains informationnels, ou encore tous ceux dont la fréquence est au dessus d'un certain seuil, etc. Autant de modules complètement indépendants pour configurer l'affichage du lexique. On pourrait bien évidemment rajouter d'autres modules sans que cela affecte ceux déjà présents. Par ailleurs il est permis que l'utilisateur puisse lemmatiser le lexique des classes comme il peut en retirer les mots fonctionnels. L'utilisateur peut appliquer l'opération de lemmatisation à l'ensemble des lexiques de toutes les classes ou seulement au lexique d'une seule classe, ceci en fonction de contraintes de temps. Il est à retenir cependant que la lemmatisation et la suppression des mots fonctionnels n'interviennent que pour améliorer l'aspect des résultats et n'interviennent nullement avant la classification à proprement parler. Toutes ces configurations du lexique sont à même d'aider l'utilisateur à proposer son interprétation des résultats.

Dépendant des paramètres choisis une chaîne de traitement spécifique se crée. Comme nous le verrons à l'aide des exemples de la prochaine section, les résultats de GRAMEXCO peuvent servir à plus d'une finalité. Ainsi, nous pouvons :

- déterminer le contenu lexical des segments similaires, et ainsi connaître le thème principal de ces segments ;
- déterminer l'acception et la signification d'un mot de par les mots qui lui sont associés dans une classe donnée ; et
- construire des classes de mots formés à partir d'un radical commun comme pour l'exemple avec informatiser, informatisation, et informatique.

Au delà de ces trois exemples d'objectif possibles, en rajoutant d'autres modules en aval de la classification, nous estimons qu'il est possible de réaliser un certain nombre de tâches telles que l'indexation, la catégorisation, etc.

En rajoutant en amont un module de requête à partir du WEB, et en remplaçant le module qui segmente le texte par un autre qui considère la page WEB un segment, nous avons pu greffer GRAMEXCO à l'INTERNET. La nouvelle chaîne de traitement s'appelle ainsi eGRAMEXCO. Elle permet entre autre d'affiner les recherches sur le WEB. Nous rappelons à cet effet que la toile dans sa configuration actuelle, étant donnée l'exploitation « mercantile » dont fait objet le WEB, un moteur de recherche a de fortes chances de retourner comme

premiers résultats d'une requête avec comme mot clé *Mozart*, un site dédié à la vente de t-shirts à l'effigie d'Amadeus et non au célèbre compositeur (test fait avec les moteurs de recherche Netscape et Altavista). Ce phénomène ne fera que s'accentuer du fait des gains financiers engendrés par une telle forme de publicité ainsi que de la chute en bourse du secteur technologique Internet. Avec eGRAMEXCO, Le résultat serait une classification des sites selon leur contenu. L'usager ne fera qu'accéder aux sites dont le contenu l'intéressera effectivement.

3.3. Évaluations de GRAMEXCO et Commentaires :

Nous avons mené deux évaluations principales (et deux complémentaires). La première voulait montrer le comportement d'une classification numérique fondée sur les n-grams de caractères. Elle a été réalisée sur un corpus formé d'une cinquantaine de pages (format ASCII) construit à partir d'extraits de documents trouvés sur le web. Ces documents couvrent divers domaines et permettent une hétérogénéité du contenu du corpus et, par conséquent, une meilleure compréhension des résultats de la classification. La deuxième évaluation avait pour but d'expliquer pourquoi la classification avec les n-grams pouvait être aussi performante sinon plus performante qu'une "classification + lemmatisation". Cette évaluation a été réalisée sur un texte de deux pages. Sa finalité n'en exigeait pas plus pour construire des classes de mots ayant un même radical.

Pour les opérations préliminaires de la **première évaluation principale**, soit la segmentation et l'extraction des n-grams, nous avons opté pour les paramètres suivants : 10 (phrases) pour déterminer la taille d'un segment et 4 (caractères) pour déterminer la taille des n-grams[2]. De plus, à l'aide des paramètres de GRAMEXCO, nous avons considéré les lettres majuscules identiques aux lettres minuscules et nous avons remplacé les caractères non alphanumériques et les chiffres par des espaces. Nous avons ainsi récupéré 174 segments et 4 857 quadri-grams, après un "ménage" de la liste des n-grams qui a consisté à supprimer les n-grams contenant un ou plusieurs espaces et les n-grams ayant une fréquence égale à 1. La classification elle-même, au moyen du réseau de neurones ART avec un paramètre de vigilance de 0.1, donne lieu à la production de 100 classes de segments présentants des similarités. Examinons maintenant quelques résultats :

- La classe 100 regroupe les segments 137 et 157. Le lexique de cette classe formé de l'intersection des lexiques des deux segments est constitué par : {bourse, francs, marchés, millions, mobile, pdg, prix}. On constate, au regard de ce lexique, que le mot francs désigne la monnaie française et n'a aucun rapport avec la franchise ou avec les fameuses tribus "les francs". Ce même lexique nous renseigne également sur le thème commun que se partagent les segments 137 et 157, en l'occurrence le domaine financier.
- La classe 54 regroupe les segments 141 et 143. L'intersection des lexiques des segments de la classe 54 est formée de : {appel, cour, décidé, juge}. Ainsi pour le mot cour, une seule signification est possible au regard des mots qui l'accompagnent : cour de justice. On écarte aisément les sens suivants : la cour qu'on fait à une demoiselle, la cour de récréation, ou encore les toilettes des Belges. Le thème de la classe 54 est par ailleurs bien identifié en ce sens qu'il s'agit de segments dont le contenu traite d'affaires judiciaires.

1.1.1.1.1.1.1.1.1.

[2] Selon Damashek (1995), les quadri-grams donneraient les meilleurs résultats pour l'anglais. Lelu *et al.* (1998) semblent confirmer cela pour le français.

- La classe 98 regroupe les segments 71 et 73. Le lexique issu de l'intersection des lexiques des deux segments est formé des mots : {culture, économiques, eurasistes, matérialiste, occident}. Dans ce contexte, le terme culture ne peut signifier que culture économique, et son utilisation n'est pas pour introduire une quelconque notion d'agriculture. Ce qui se confirme d'ailleurs avec le mot occident qui est utilisé ici dans le sens bloc géopolitique et non "là où se couche le soleil". Le thème de la classe 98 traite sans conteste d'options économiques ce que nous pouvons d'ailleurs vérifier au travers de la lecture des segments 71 et 73.
- La classe 64 regroupe les segments 166 et 167. Le lexique qu'on retiendra pour cette classe est formé de tous les mots dont la fréquence dans les segments 166 et 167 est supérieure ou égale à 2, en l'occurrence les mots : {chance, dernière, dire, match, stade, supporters, vélodrome}. Le mot stade, du fait particulièrement de la présence des mots match, supporters et vélodrome, est compris comme étant un stade de football. Par ailleurs, pour un public averti qui sait que le vélodrome est le stade de Marseille, on comprend aisément que les deux segments 166 et 167 traite des supporters de l'Olympique Marseillais.
- La classe 13 regroupe les segments 32, 35, 41 et 48. Le lexique de cette classe formé de l'intersection des lexiques des quatre segments est constitué du seul mot : russe. Celui-ci est suffisant pour nous permettre de conclure que le thème partagé par les quatre segments se rapporte à la Russie. L'union des lexiques, formée entre autres des mots : conservateur, socialisme, marxiste, conservateur, révolutionnaire, Dostoievski, doctrine, impérial, slavophile, etc., nous permet de préciser que le thème de la classe 13 est dédié aux slavophiles et à la culture politique russe du 19ième siècle. Une remarque s'impose : on imagine mal comment une classification fondée sur les mots aurait pu arriver à regrouper les segments 32, 35, 41 et 48 dans la même classe sans avoir recours à la lemmatisation étant donné que le seul mot commun est russe. Reste que la lemmatisation est relativement coûteuse en temps d'exécution et est une opération spécifique à chaque langue. Nous évitons ces inconvénients en utilisant les n-grams de caractères.

Les raisons d'aussi bonnes performances nous les retrouvons dans les résultats de la **deuxième évaluation principale**. En effet, celle-ci consistait à passer un texte de deux pages formé d'extraits d'un corpus sur les biotechnologies (utilisé dans Biskri et Delisle, 2000) par une classification basée sur les n-grams avec, comme paramètre, n=4 et la taille du segment ramenée à un mot seulement. Ainsi les segments regroupés dans une même classe seraient constitués des mots ayant des points communs, en particulier un radical commun et, donc, référant à une notion commune. Cette évaluation a permis de construire l'échantillon de classes suivantes :

Classe 101 : {survécu, survie}

Classe 102 : {utilisée, outil}

Classe 110 : {congelé, décongelé, congelés, congélateur}

Classe 112 : {simple, simplifier, simplifiée}

Classe 162 : {avenir, devenir}

Classe 4 : {principale, principalement}

Classe 48 : {optimisées, optimum}

Classe 60 : {cellules, cellulaire}

Classe 65 : {collecte, collectifs}

Classe 7 : {transfert, transférables, transférés, pénétrant, transferts, retransfert}

Classe 81 : {glycol, glycérol}

Classe 88 : {déshydratées, déshydratation}

Si nous prenons la classe 110 par exemple, nous nous apercevons que non seulement congelé et congelés sont regroupés, ce qu'aurait d'ailleurs fait une lemmatisation standard, mais en plus la classification leur associe décongelé et congélateur. En somme la classe 110 regroupe tout ce qui se rapporte à la notion de congélation. Il en est de même pour les autres classes qui chacune regroupe des mots partageant des notions communes. Ainsi la classe 101 porte sur la notion de survie, la classe 102 sur la notion d'outils utiles, la classe 112 sur la simplicité, la classe 48 sur l'optimalité, la classe 65 sur la notion de cellule, la classe 65 sur celle de la collection, la classe 7 sur la notion de transfert, la classe 81 sur un concept chimique particulier, et la classe 88 sur la notion de déshydratation.

Nous avons aussi effectué une **troisième évaluation**, complémentaire aux deux premières. Elle a consisté particulièrement à comparer les résultats de GRAMEXCO lors de la première évaluation à des résultats obtenus avec NUMEXCO (Biskri et Delisle,1999 ; Meunier *et al*., 1997) — NUMEXCO est une autre chaîne de traitement , contrairement à GRAMEXCO, considère comme unité d'information le mot (et non les n-grams de caractères) et donc utilise un module d'extraction des unités d'information différents de celui GRAMEXCO utilise. Nous avons soumis à NUMEXCO le même texte ASCII que lors de notre seconde évaluation et ce, avec le même paramètre de segmentation. Nous avons obtenu un lexique de 4 884 mots, après lemmatisation et suppression des mots fonctionnels. La suppression des hapax (mots dont la fréquence est égale à 1) aurait diminué le lexique à 1 755 unités d'information. Notons cependant que la suppression des hapax renvoie en quelque sorte à une suppression de n-grams dont la fréquence pourrait dépasser 1, ce qui n'était pas le cas dans notre première évaluation où nous n'avons supprimé que les n-grams dont la fréquence était égale à 1 — ce facteur donne lieu à une comparaison biaisée. Ceci dit, il est important de souligner que pour un texte ne dépassant pas les 200 pages environ, la taille du lexique et le nombre de n-grams ne diffèrent pas de beaucoup. Pour certains textes, le nombre de n-grams peut dépasser la taille du lexique. Ce qui pourrait inciter l'usager à choisir la chaîne de traitement NUMEXCO étant donnée que la taille du lexique est des fois un handicap majeur dans les opérations de classification. Cependant, dès lors que la taille du texte dépasse les 200 pages, voire les 300 pages, la taille du lexique tend à augmenter alors que le nombre de n-grams se stabilise. L'usager das ce cas préférera probablement GRAMEXCO. Sur le plan de la classification à proprement parler, les classes que nous sommes arrivés à construire avec GRAMEXCO ont été impossibles à reproduire avec NUMEXCO et ce, en raison du peu de mots communs après lemmatisation se trouvant dans les différents segments des classes.

Finalement, dans le cadre d'une **dernière évaluation** complémentaire, qui consiste en réalité en un retour à la question du multilinguisme avec la soumission à GRAMEXCO d'un corpus constitué d'un texte anglais de 20 pages portant sur des sujets d'actualité. Ce texte a été récupéré à partir du site de CNN. Il est composé d'articles de presse. De par cette dernière évaluation nous voulons prouver que des résultats semblables à ceux obtenus avec le français peuvent être obtenus avec l'anglais.

La première partie de notre dernière évaluation consistait à soumettre l'ensemble du corpus anglais à GRAMEXCO avec les paramètres suivants : le segment =10 phrases, le n-gram = quadri-gram, suppression des n-grams ayant une fréquence de 1, suppression des n-grams contenant des espaces, classification au moyen de ART avec un paramètre de vigilance de 0.1.

Les résultats : 3556 quadri-grams, 102 segments et 54 classes dont nous présentons un échantillon ci-après.

- La classe 16 regroupe les segments 33 et 34. Le lexique de cette classe formé de l'intersection des lexiques des deux segments est constitué par : {station, shuttle, space, russian, nasa, launch, dock }. On constate, au regard de ce lexique, que le mot space désigne l'espace dans son sens cosmique et non un intervalle. Il en est de même pour shuttle qui désigne dans le cas présent une navette spatiale et non le mouvement alternatif (shuttle movement). Ce même lexique nous renseigne également sur le thème commun que se partagent les segments 33 et 34, en l'occurrence la conquête spatiale.
- La classe 2 regroupe les segments 2, 4 et 5. Le lexique de cette classe 54 est formée de : {court, investigation, israeli, sharon}. Ainsi pour le mot court, une seule signification est possible au regard des mots qui l'accompagnent : cour de justice. On écarte aisément les sens suivants : ruelle, ou encore le verbe courtiser.
- La classe 24 regroupe les segments 53, 54 et 55. Le lexique est formé des mots : {hospitals, patient, Hollebeek, project, computing, data, cancer, breast, built, grid}. Dans ce contexte, le terme patient ne peut signifier que malade, et son utilisation n'est pas pour introduire une quelconque notion de patience ou d'endurance. Le thème de la classe 24 traite sans conteste d'un projet médical en rapport avec le cancer du sein.
- La classe 4 regroupe les segments 10 et 9. Le lexique est formé des mots : {Afghanistan, government, members, security}. Le mot members est compris comme étant un membre (comme ici d'un gouvernement) et non comme un organe anatomique. On comprend par ailleurs aisément que les deux segments 10 et 9 traite de la sécurité des membres du gouvernement afghan.
- La classe 44 regroupe les segments 98, 99, 100, 101, 102. Le lexique de cette classe formé de l'intersection des lexiques des cinq segments est constitué de {central, carat, diamonds, model, platinum, plain, weighing, head, hoop}. Pour un américain le mot diamonds correspondra à une pierre précieuse et non à un terrain de base-ball et que d'autre part la classe représente un contenu textuel qui traite de bijoux.

La deuxième partie de notre dernière évaluation consistait à limiter la taille du segment à un mot tout en gardant la taille du n-gram à 4. Cette évaluation est utile pour savoir si GRAMEXCO est assez puissant pour reconnaître des mots anglais ayant le même radical, comme nous avons pu le montrer pour le français. Nous avons obtenu les classes suivantes :

Classe 85 : {peace, peacekeepers, peecekeeping}

Classe 97 : {accused, accusations}

Classe 16 : {Israel, israeli, israelis}

Classe 116 : {Lebanon, lebanese}

Classe 107 : {inquiries, required, inquiry}

Classe 101 : {law, lawyers}

Classe 130 : {minor, minorities, minority}

Classe 133 : {civilians, civilized}

Classe 110 : {allegations, alleged}

Classe 231 : {city, citizen}

Classe 52 : {Belgium, belgian, belgians}

Classe 14 : {thursdays, wednesday, tuesday}

Classe 18 : {reports, reporters}

Classe 212 : {imprisonment, prison, prisoners, prisons}

Classe 25 : {agents, agency}

Classe 35 : {congress, congressional, congressmen, agressive, progress}

Classe 41 : {face, surface}

Classe 6 : {north, northern}

Classe 60 : {prosecute, prosecuted, prosecutor, security}

Cette dernière évaluation ne fait que confirmer les deux premières évaluations faites pour le français. Il n'y aura pas lieu donc de la commenter. Toutefois elle confirme toute l'importance de la flexibilité de SATIM qui doit permettre la création aussi facile que possible de plusieurs chaînes de traitements (comme nous l'avons montré par un jeu de paramètres) pour répondre soit à un objectif précis (par exemple le multilinguisme comme nous l'avons vu), soit à apporter une réponse à une ou plusieurs contraintes (par exemple la taille du lexique).

Il reste certainement un énorme travail à faire.

4. Conclusion

Notre travail semble très concluant quant à l'importance d'une plate-forme comme SATIM. Nous avons donné des exemples et des évaluations par rapport à des données textuelles mais notre conviction reste qu'il est possible de construire des chaînes de traitement pour d'autres sources d'information que des sources textuelles. L'analyse multidimensionnelle est assez générale pour qu'on puisse sérieusement envisager de rajouter des modules qui extraient des unités d'information pour le son, l'image, etc. et des modules qui permettent de définir des segments pour ces mêmes sources d'information.

Au delà de considération purement technique, SATIM laisse entrevoir une nouvelle technologie « pièces détachées » où un expert en ingénierie de l'information pourra choisir des modules et construire une chaîne de traitement tel un enfant construisant un « petit quelque chose » avec son jeu de LEGO. Une multitude de chaînes de traitement seront possibles. Il sera même permis dans les prochaines versions de SATIM de leur donner un nom et de les stocker en mémoire pour une réutilisation. Ce qui évitera à un usager de refaire manuellement toutes les étapes de choix des paramètres et ainsi faire son traitement offline. Pour ce faire nous

aurons besoin d'imprimer dans une base de données par exemple la combinaison des modules dans la chaîne de traitement. Un moyen élégant pour représenter cette combinaison pourrait être la logique combinatoire (Biskri & Desclés, 1997). Cette dernière pourrait également nous aider à détecter l'ensemble des combinaisons impossibles. Mais ce dernier point reste du domaine des perspectives.

Enfin, SATIM nous laisse envisager une possible coopération avec d'autres chercheurs qui aboutirait probablement vers une sorte de communauté scientifique inspirée du modèle LINUX sans pour autant être totalement **open source** comme l'est LINUX.

Références

Balpe, J.P., Lelu, A. Papy, F. (1996), *Techniques avancées pour l'hypertexte*. Paris, Hermes.

Benhadid, I., Meunier, J.G., Hamidi, S., Remaki, Z., Nyongwa, M. (1998), "Étude Expérimentale Comparative des Méthodes Statistiques pour la Classification des Données Textuelles", *Actes de JADT-98*, Nice, France.

Biskri, I., Delisle, S. (2000), "User-Relevant Access To Textual Information Through Flexible Identification Of Terms: A Semi-Automatic Method And Software Based On A Combination Of N-Grams And Surface Linguistic Filters", *Actes de RIAO-2000*, Paris, France, 1059-1068.

Biskri, I., Delisle, S., (1999), "Un Modèle Hybride pour le Textual Data Mining : Un Mariage de Raison entre le Numérique et le Linguistique", *Actes de TALN-99*, Cargèse, France, 55-64.

Biskri, I., Desclés, J.P., (1997), "Applicative and Combinatory Categorial Grammar (from syntax to functional semantics)", dans Recent Advances in Natural Language Processing (selected Papers of RANLP 95) Ed. Ruslan Mitkov & Nicolas Nicolov. John Benjamins Publishing Company, Numéro 136, pages 71-84.

Damashek, M., (1995), "Gauging Similarity with n-Grams : Language-Independent Categorization Of Text", *Science*, 267, 843-848.

Greffenstette, (1995), "Comparing Two Language Identification Schemes", *Actes de JADT-95*, 85-96

Halleb M., Lelu A., (1998), "Hypertextualisation Automatique Multilingue à Partir des Fréquences de n-Grammes", *Actes de JADT-98*, Nice, France.

Lelu A., Halleb M. , Delprat B. (1998), "Recherche d'information et Cartographie dans des Corpus Textuels à Partir des Fréquences de n-Grammes", *Actes de JADT-98*, Nice, France.

Manning, C.D., Schütze, H., (1999), *Foundations of Statistical Natural Language Processing*, MIT Press.

Mayfield, J., Mcnamee, P., (1998), "Indexing Using both n-Grams and Words", *NIST Special Publication 500-242 : TREC 7*, 419-424.

Meunier, J.G., Biskri, I., Nault, G., Nyongwa, M. (1997), "Aladin et le Traitement Connexionniste de l'Analyse Terminologique", *Actes de RIAO-97*, Montréal, Canada, 661-664.

Remaki, L., Meunier, J.G., (2000), "Un modèle HMM pour la détection des mots composés dans un corpus textuel" Actes de JADT-2000, Lausanne, Suisse.

Rialle, V., Meunier, J.-G., Oussedik, S., Biskri, I., Nault, G., 1998, "Application de l'Algorithmique Génétique à l'Analyse Terminologique", Acte du colloque international JADT 98, Nice, France.

Turenne, N. (2000), *Apprentissage statistique pour l'extraction de concepts à partir de textes (Application au filtrage d'informations textuelles),* thèse de doctorat en informatique, Université Louis-Pasteur, Strasbourg, France.