Coopération multiniveau d'approches non supervisées et supervisées pour la détection des ruptures thématiques dans les discours présidentiels français

Alain Lelu, Martine Cadot, Sylvain Aubin

Coopération multiniveau d'approches non-supervisées et supervisées pour la détection des ruptures thématiques dans les discours présidentiels français.

Alain Lelu*, Martine Cadot**, Sylvain Aubin***

*Université de Franche-Comté / LASELDI
30 r. Mégevand, 25 030 Besançon cedex
alain.lelu@univ-fcomte.fr
**UHP Nancy / LORIA
Bât. C - BP 239 - 54506 Vandœuvre lès Nancy cedex
martine.cadot@loria.fr
http://www.loria.fr/~cadot
***Diatopie
27 bd. Saint-Martin 75003 Paris
sylvain.aubin@diatopie.com
http://www.diatopie.com

Résumé. Notre réponse à la tâche de détection des ruptures thématiques repose sur la construction d'un petit nombre d'indicateurs numériques à valeur croissante avec la probabilité pour une phrase d'être un début de paragraphe thématique, chacun traduisant un point de vue différent. Deux indicateurs expriment directement ce qu'on peut induire des débuts de phrases. Une phase de forte réduction de dimensions du problème, non supervisée, était un préalable pour les autres points de vues, que ce soit le rhétorique, où chaque phrase a été réduite à un vecteur à 100 dimensions, ou le sémantique, où chaque paragraphe a été réduit à 200 dimensions. Ce dernier cas a posé un difficile problème d'apprentissage de données complexes multiniveau, auquel nous avons apporté un début de réponse. L'apprentissage final par règles de décision de nos 5 indicateurs, perfectible, nous a conduit à des performances honorables par rapport aux autres équipes.

1 Introduction

La tâche demandée aux participants au défi DEFT'06 (DEFT 2006) était de retrouver un découpage prédéfini au sein de trois ensembles de textes, chacun n'étant connu que par la séquence indifférenciée de ses phrases. Dans le cas du corpus des discours présidentiels français, ce découpage en segments thématiquement homogènes a été établi manuellement par un service de documentation, et les thèmes vraisemblablement caractérisés par des termes pris dans un vocabulaire prédéfini, non fourni aux participants. Pour le corpus « Ouvrage scientifique », ce découpage était celui de la structure en chapitres et sections, et dans le cas du corpus Eur-Lex celui constitué par les lois européennes.

La possibilité de réinvestir des travaux antérieurs, ainsi que le manque de temps et de moyens humains, nous ont poussés à ne traiter que le corpus des discours présidentiels

Les difficultés posées par ce corpus ne venaient pas tant de sa taille, proche de la limite de nos outils matériels et logiciels, sans la dépasser, que de son hétérogénéité :

. Hétérogénéité de forme : la partie Giscard d'Estaing a été saisie en lettres capitales désaccentuées, et comporte des traits conventionnels (traits de soulignement, tirets) placés dans la louable intention de distinguer l'usage stéréotypé de certains mots, comme *plan* (dans *sur le* ***plan*** *de*…vs. *commissariat au* ***Plan***), mais cette pratique coûteuse n'a pas été poursuivie dans les autres parties du corpus.

. Hétérogénéité de fond : le corpus comprend, en plus des discours proprement dits, des interviews, débats, adresses et réponses de personnalités étrangères (Tony Blair, Yasser Arafat…).

. Hétérogénéité du découpage en paragraphes[1] thématiques : si le nombre moyen de phrases par paragraphe est à peu près constant dans les parties Giscard et Mitterrand (autour de 12 et 14 respectivement, il augmente considérablement dans la partie Chirac (de 20 dans le premier tiers du corpus d'apprentissage à plus de 60 dans le dernier tiers ! (cf. Fig. 1), traduisant sans doute un changement de politique éditoriale (ou un changement de personne ?) dans le service de documentation concerné.

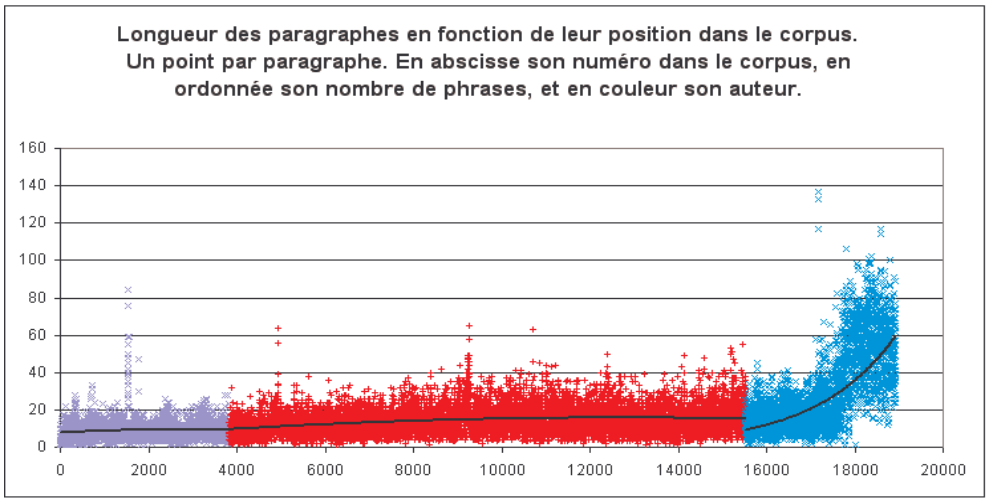

FIG. 1 – *Longueur des paragraphes. À gauche (en violet): VGE, milieu (rouge): Mitterrand, droite (bleu): Chirac*

Mais le défi le plus intéressant venait de la nature complexe des objets traités : contrairement aux tâches classiques d'apprentissage supervisé où il s'agit d'attribuer à un ensemble d'objets des catégories prédéfinies, nos objets-phrases (connus) ne sont que les composants d'objets de niveau supérieur (les « paragraphes » thématiques) dont il est

[1] Convention de vocabulaire : nous appelons paragraphe l'unité textuelle considérée comme homogène par les documentalistes concernés. Nous avons doté ses phrases d'un code de position : 1 pour la phrase de début, -1 pour celle de fin, 0 pour les autres, mais nous n'avons pas exploité le code -1 dans le travail présenté ici.

demandé d'établir les frontières, alors que le nombre et la nature des catégories thématiques sont cachés, même dans l'ensemble d'apprentissage.

2 Principes mis en œuvre

. *Réduire les dimensions du problème*, éliminer le bruit au maximum, par des approches non supervisées (Cornuéjols et Miclet, 2002). Ceci dit, notre expérience nous montre que ce processus provoque des « déchets », à savoir des phrases qui se retrouvent sans termes d'indexation, ou avec des termes très peu pris en compte dans l'analyse, donc avec des valeurs nulles sur toutes les dimensions. Un compromis est donc à trouver entre la réduction du nombre de dimensions et le nombre de phrases « oubliées ».

. *Mettre en œuvre plusieurs angles d'attaque* = plusieurs "vues", les plus orthogonales possibles, indépendamment, sur le problème, avant d'en organiser la coopération dans la phase d'apprentissage finale – plusieurs juges indépendants valent mieux qu'un seul, ou que plusieurs juges apparentés, l'idéal étant que dans le tout final les erreurs des parties se compensent mutuellement. Ici, trois vues principales ont été retenues : 1) les débuts de phrases (aspects formels de méta-informations, et « attaques » de phrases), 2) les traits rhétoriques présents dans les phrases, quelle que soit leur localisation, 3) la thématique (sémantique) traitée dans chaque paragraphe.

. *Prendre en compte la nature intrinsèquement multiniveau du problème* :

- On dispose d'une séquence indifférenciée de phrases,
- ces phrases s'enchaînent pour former des unités thématiques textuelles, les paragraphes ; les limites de ces unités sont connues dans le corpus d'apprentissage, mais le nombre et la nature des thèmes restent inconnus,
- ces paragraphes ont un début et une fin : il faut opérer un retour au niveau phrase pour déterminer les phrases de début, corps de texte et fin.

. *Mettre en œuvre* partout où cela est possible *des optimisations dans l'espace des paramètres* : ceci intervient dans la construction des indicateurs numériques de début de paragraphe, dans le choix des seuils de coupure sur variables quantitatives pour les règles de décision, dans les passages entre niveaux de données (agrégation / désagrégation entre phrases et paragraphes thématiques). Nous avons ici la chance de disposer d'un critère numérique clair d'évaluation des choix et des procédures, utilisable comme fonction objectif à maximiser. Il reste qu'une telle optimisation n'a pas été possible là où la taille de l'espace des paramètres est excessive, en particulier dans la phase non supervisée d'analyse textuelle : si en théorie rien ne s'oppose au choix « informé par la fonction objectif » des termes pertinents (sélection de variables) ni à celui des paramètres de nombre de classes demandées et d'initialisation de ces classes, la combinatoire à mettre en œuvre rend ceci impossible, et on s'est contenté pour cette partie de décisions de bon sens et d'expérience.

3 Outils et environnements de travail utilisés

L'approche non supervisée a été réalisée sous l'environnement d'indexation et de *clustering* textuel NeuroNav (Lelu et Aubin 2001), tant pour les aspects rhétoriques que

sémantiques. Nous décrirons plus loin les fonctions de NeuroNav qui nous ont été utiles pour mener à bien ces tâches.

L'approche supervisée a fait l'objet d'une programmation directe en Python, langage de script souvent utilisé en fouille de textes (Ziadé 2006). Diverses tâches complémentaires, comme les transformations de formats de fichiers, certains traitements auxiliaires ou exploratoires, ont été réalisés en Perl, Scilab ou sous Excel.

4 Réalisation des trois angles d'attaque

La « matière textuelle » brute se caractérise par le très grand nombre d'attributs susceptibles de la décrire, à la limite des capacités actuelles de l'informatique (plus d'un million de formes différentes trouvées sur le corpus d'apprentissage de 300 000 phrases, hapax et formes composées compris !) , et par la diversité des points de vues que l'on est susceptible d'adopter pour les regrouper et les sélectionner (regroupements syntaxiques en lemmes, en expressions composées, ou filtrages sémantiques, syntaxiques ou statistiques, ...).

Pour nous, la sélection directe d'attributs en rapport avec la tâche demandée – sauf à considérer un petit sous-ensemble de ceux-ci, comme les débuts de phrases – est 1) non traitable computationnellement, 2) inadaptée au caractère multiniveau de la tâche demandée. Il faut donc une étape préalable de « débruitage » de cette masse d'information brute, d'autant que celle-ci peut aboutir à donner au problème un nombre de dimensions raisonnable.

Nous avons privilégié deux points de vue dans nos opérations de réduction de dimensions :

. un point de vue rhétorique centré sur la mise en évidence de traits de langage récurrents, de formules toutes faites, de tics, d'éléments de « langue de bois » ou d'expressions convenues ou de politesse, quelle que soit leur position dans la phrase.

. un point de vue sémantique centré sur l'extraction de contextes homogènes du point de vue du contenu, par exemple ce qui a trait à la politique étrangère, ou à l'environnement – contextes qui dépassent nécessairement les limites de la phrase individuelle et sont construits séquentiellement par l'accumulation de ces phrases ; ceci afin de détecter les transitions entre contextes.

Nous avons aussi choisi d'utiliser directement les chaînes de caractères, dans la limite d'une vingtaine de premiers caractères de chaque phrase, car ceux-ci contiennent une information importante et non réductible aux deux points de vue précédents :

. des méta-informations introduites par les transcripteurs et documentalistes (« *(inaudible)* », « *Paul Amar :* », …),

. des « attaques » de phrases (« *En ce qui concerne…* », « *C'est pour cela que…* », …), des enchaînements pour capter l'attention du public et structurer le propos.

Hypothèses et vérifications

Nous explicitons ici nos deux principales hypothèses :

H1 : *Il existe une similitude entre phrases de même position dans les « paragraphes » (unités textuelles homogènes déterminées par le service de documentation source).*

Cette similitude peut provenir d'éléments purement formels de méta-information (ex. : *(Inaudible)*), aussi bien que de tics rhétoriques, ou de la présence plus ou moins importante d'éléments sémantiques communs (mots « pleins » établissant le thème du discours).

H2 : *Il existe une similitude entre les phrases dans un même paragraphe.*
En d'autres termes : elles parlent du même sujet.
La difficulté se trouve dans la procédure d'agrégation/désagrégation entre un niveau connu dans les ensembles d'apprentissage et de test (la phrase) pour lequel la notion de thème sémantique revêt peu de signification individuellement – c'est le contexte gauche de la phrase, au sein du paragraphe, qui lui donne son sens – et un niveau supérieur, le paragraphe, pour lequel cette notion est pleinement opérationnelle, mais qui n'est pas connu dans le corpus de test.

Notre hypothèse secondaire H3 est qu'il est possible de déterminer pour chaque phrase une attribution d'auteur grossière[2] (cf. plus haut section 1). Nous avons ainsi créé une nouvelle variable susceptible d'être prise en compte dans l'apprentissage.

Rendre opérationnelles ces hypothèses revient pour nous à créer un indicateur numérique unique, pour chaque point de vue, qui soit une fonction monotone de la probabilité d'une phrase d'être début de paragraphe. Par exemple, le plus simple indicateur que nous ayons construit est le nombre de mots sémantiquement significatifs (cf. section 4.4.3 plus bas) de chaque phrase : la valeur 0 correspond à la moindre probabilité constatée d'être début de phrase, la valeur 1 à une probabilité légèrement plus élevée, etc. Le paragraphe suivant donne le principe de construction d'un indicateur à partir d'une information qualitative (autrement dit : à partir d'une variable nominale à modalités exclusives).

L'hypothèse H1 ne fait intervenir que des informations locales à chaque phrase, accessibles aussi bien dans le corpus d'apprentissage (sa position dans le paragraphe, son contenu) que de test (son contenu). Notre construction d'indice la concernant, dans le cas qualitatif, fait le plus souvent intervenir l'« auteur » présumé, suivant le schéma :

- la connaissance de l'auteur présumé permet de rechercher la répartition de cette variable chez cet auteur selon les positions de la phrase dans le paragraphe[3].
- le rapport entre fréquence en position de début et fréquence en position autre que de début et fin donne la valeur de l'indicateur (en cas d'inexistence de cette répartition, une valeur par défaut est attribuée, ici 0,5).

L'hypothèse H2 est beaucoup plus difficile à rendre opérationnelle :
Le cadre pour formaliser cette hypothèse est clair : on établit pour chaque phrase une valeur 'val_d' de ressemblance sémantique avec les phrases qui la suivent et une valeur de

[2] La coupure VGE-Mitterrand à la fin des lettres capitales, la coupure Mitterrand-Chirac en amont de la première occurrence de « Président Chirac ». Dans le corpus de test, nous avons placé la deuxième coupure au début du débat télévisé Chirac-Mitterrand (dans le sens d'un léger rééquilibrage quantitatif en faveur du corpus Chirac). Nous avons donc étiqueté « Chirac » des phrases de Mitterrand - mais aussi de Yasser Arafat, Jian Zemin…- et « Mitterrand » des phrases de Tony Blair !

[3] La connaissance de ces répartitions pour la phrase précédente et la suivante permettrait d'utiliser l'information de séquentialité des phrases pour renforcer la fiabilité de l'indicateur, si ces indications sont cohérentes, ou la diminuer sinon, mais nous ne l'avons pas fait par manque de temps.

ressemblance ‘val_g’ avec celles qui la précèdent. La différence val_d-val_g est proche de 0 si on est au milieu du paragraphe car alors la phrase considérée ressemble autant à celles de gauche qu’à celles de droite ; et au fur et à mesure qu’on s’approche de la fin d’un paragraphe, la phrase a tendance à ressembler de moins en moins à celles de droite, et de plus en plus à celles de gauche. Ainsi la différence val_d-val_g est négative et diminue de plus en plus quand on va du centre du paragraphe à la fin du paragraphe. C’est là qu’elle atteint sa plus grande valeur négative. Inversement si on remonte dans le paragraphe en prenant une phrase à gauche de celle du milieu, il y a plus de phrases à sa droite qui lui ressemblent qu’à sa gauche, et la différence val_d-val_g est alors positive et atteint son maximum quand la phrase est prise au début du texte[4]. Par cette définition, on crée une fenêtre de lissage qui permet de prendre en compte le fait qu’un paragraphe peut contenir des phrases ne laissant apparaître aucun de ses thèmes sémantiques privilégiés (notamment certaines phrases n’ont aucun mot porteur de sens). La largeur de la fenêtre de lissage est un paramètre à déterminer pour limiter le plus possible les variations désordonnées de cette différence. On peut aussi jouer sur sa forme par l’utilisation de poids permettant de donner moins d’importance aux phrases plus éloignées.

Mais nos plus gros efforts ont porté sur la définition d’une similarité entre phrases, que notre traitement non-supervisé va caractériser à la fois par des éléments symboliques et numériques (cf. section 4.4.5 plus bas).

Pour chaque hypothèse nous avons défini un ou deux indicateurs numériques fonctions de nombreux paramètres quantitatifs ou non, comme des valeurs de seuil, des tailles de listes, des variantes de calcul, … La vérification d’une hypothèse consiste 1) à déterminer un jeu optimal de paramètres pour un indicateur, 2) à trouver par balayage un seuil de coupure sur cet indicateur, tel que la détection de débuts de phrases présumés au dessus de ce seuil soit la plus conforme possible aux débuts de phrases réels, au sens du Fscore[5] strict ou flou. Ce qui évite au maximum des décisions arbitraires ou subjectives de « designer » ou d’utilisateur d’algorithmes, et constitue un type de méta-apprentissage (Dussart et Petit 2005).

Nous avons donc fait varier les paramètres sur des grilles d’étendues et de pas prédéterminés, et calculé pour chaque combinaison les Fscore obtenus sur nos ensembles d’apprentissage, d’ajustement et de test[6], résultats conservés dans un journal utilisé pour choisir quelques jeux de paramètres donnant des valeurs de Fscore stables pour des petites variations de leurs valeurs sur le premier ensemble, et donnant pour ces variations des valeurs de Fscore dont le minimum était le plus élevé possible pour les deux autres ensembles (stratégie du risque minimal).

Les débuts de phrases : aspects formels et « attaques ».

Comme présenté plus haut, nous avons utilisé deux indicateurs voisins, non orthogonaux (coefficient de corrélation linéaire de Bravais-Pearson de 0,758 = cos 41°), calculés sur le

[4] Une différentiation supplémentaire de ces différences (δ_i=valeur pour phrase i – valeur pour phrase i-1) nous permettra d’accentuer encore le marquage des limites de paragraphe.

[5] Fscore : mesure globale de qualité d’apprentissage, moyenne harmonique des taux de rappel et de précision.

[6] Un seul découpage du corpus en 3 parties inégales « représentatives » a été utilisé, alors qu’une validation croisée sur découpage « tournant » au sein de chaque partie de même auteur aurait été préférable, nous en sommes conscients…

principe suivant : on attribue à chaque phrase une chaîne de caractères, puis une valeur numérique fonction de la fréquence relative de cette chaîne dans les phrases de début de paragraphe chez l'auteur. Ces indicateurs sont constitués respectivement à partir :

1) des *n* premiers caractères non filtrés, donc orientés « méta-information » (détection de *(inaudible)*, etc.) ; cet indicateur s'est trouvé optimisé pour *n*=6 et un seuil de coupure de 1,0 (0,5 pour le critère flou), donnant lieu aux valeurs de Fscore suivantes :

	Fstrict	Fflou1
Appr .	0,345	0,368
Ajust.	**0,285**	0,339
Test	0,378	**0,321**

2) des premiers caractères avant premier séparateur rencontré, dans la limite de *n*, filtrés par désaccentuation, passage en minuscules, enlèvement des séparateurs et caractères non alphanumériques (orientation « attaques de phrases » : *mesdames, …*) ; l'indicateur s'est trouvé optimisé pour *n*=10, la liste de séparateurs {[ ] ['] } et un seuil de coupure de 1,0 (0,5 pour le critère flou), donnant lieu aux valeurs de Fscore suivantes :

	Fstrict	Fflou1
Appr.	0,384	0,404
Ajust	**0,279**	0,341
Test	0,357	**0,333**

Aspects rhétoriques : classification non supervisée des phrases à partir de leur indexation "plein texte".

L'originalité principale du logiciel NeuroNav est de présenter une interface ergonomique pour le contrôle et la correction du vocabulaire d'indexation issu de la phase initiale de lemmatisation des textes et extraction de termes composés, avec de nombreuses fonctions facilitant cette tâche. Ici nous avons « détourné » le module de lemmatisation de NeuroNav en utilisant un dictionnaire vide afin que les mots soient réduits aux seules chaînes de caractères (indexation en texte plein, donnant 29 000 chaînes environ, en éliminant les hapax). Par contre, la fonction d'extraction de termes composés nous a servi et nous a permis d'extraire 59 000 expressions récurrentes (hors hapax) comme *Monsieur_le_Président*, ou *comme_vous_le_savez*, … Nous avons choisi d'éliminer, en plus des expressions hapax, celles de fréquence 2, ce qui nous en a laissé 12 000. Ce choix, s'il a permis d'éviter de stocker de l'ordre d'un million de formes pour la plupart inutiles, s'est révélé un bon compromis puisqu'il y a très peu de phrases sans chaîne simple ou composée associée (0.06%), en général à la suite de coquilles comme *extraordinaireà*.

Ensuite l'extraction des thèmes rhétoriques s'est faite, de façon classique avec NeuroNav, en quelques itérations entre clustering par la méthode des K-means axiales (Lelu 1994) et corrections du paramètre K (nombre de clusters demandés) et du vocabulaire décrivant les unités textuelles, ici les phrases : 50 thèmes au départ, nombre qui s'est révélé insuffisant au regard de la variété présente dans le corpus, puis 100 thèmes ; l'examen de ces thèmes nous a convaincus que deux catégories de mots fréquents étaient à éliminer : les particules et mots grammaticaux, comme *à, a, au, ce, d, dans, de,…*, et les mots porteurs de sens comme *administration, affrontement, afghanistan,…* La consultation de la liste des termes par ordre de fréquence décroissante et les fonctions « panier de mots » et « antidictionnaire » de NeuroNav nous ont permis de constituer rapidement et appliquer un antidictionnaire de 307

termes, augmenté à l'itération suivante d'une trentaine d'autres, à la suite de quoi nous avons pu constater sur la carte d'ensemble des thèmes que ceux-ci correspondaient bien à notre attente des 100 principales classes de "tics rhétoriques" du discours présidentiel. La copie d'écran ci-après montre par exemple le contenu du thème construit autour de *non* et autres formules de négation. Nous avons ainsi extrait les 100 principales « tournures favorites » présidentielles, au milieu de leurs variantes et de leur environnement lexical, qu'il aurait été très difficile de détecter autrement, parmi des milliers d'autres.

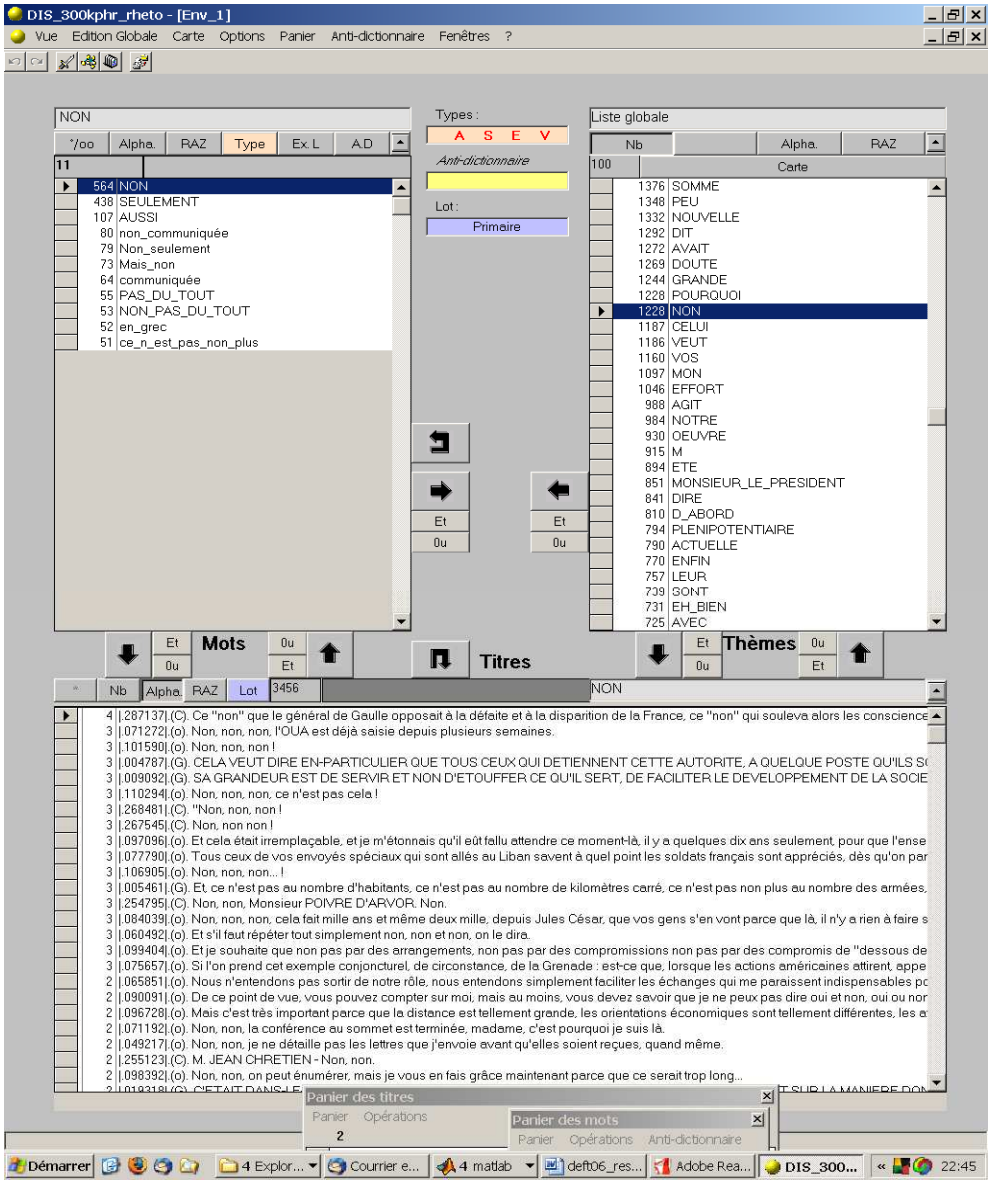

Fig. 2 – Ecran principal NeuroNav montrant le contenu du thème rhétorique « Non ». Codes auteurs : (C) Chirac, (G) Giscard d'Estaing, (o) Mitterrand.

Enfin l'indicateur « rhétorique » a été calculé pour chaque phrase à partir de son N° de thème dominant et de son auteur présumé. Par exemple, une phrase de thème dominant N° 44 (mots autour de *Monsieur_le_Président*) se voit attribuer la valeur 111/71 = 1,56 si elle se trouve dans la partie Mitterrand – sur les 1187 phrases ayant ce thème dominant, 111 sont en première position chez Mitterrand, 71 en position « corps de texte » ; sa valeur serait de 1,92 si elle était située dans la partie VGE, 0,22 dans la partie Chirac.

En ce qui concerne les résultats, l'optimum a été trouvé pour la valeur 1 du nombre de thèmes dominants, et le seuil de coupure 1,1 (1,0 pour le critère flou) :

	Fstrict	Fflou1
Appr.	0,269	0,313
Ajust.	**0,231**	0,311
Test	0,307	**0,304**

Aspects sémantiques :

Préparation des textes

Un certain toilettage des textes de VGE, en majuscules désaccentuées, a été nécessaire : suppression des traits de soulignements, des tirets « parasites », etc., pour les homogénéiser avec le reste du corpus.

Indexation automatique et contrôle du vocabulaire sous NeuroNav

L'indexation automatique en entrée de NeuroNav a été réalisée avec les options « dictionnaire désaccentué » et « suppression des hapax » en 2 heures environ pour les 303 000 phrases ; 2 600 verbes, 4 000 adjectifs, 13 000 substantifs et 37 000 expressions composées en ont résulté.

Nous avons choisi de retenir les types grammaticaux substantifs, expressions composées (avec un seuillage éliminant les fréquences <=3), et adjectifs, non sans utiliser pour ces derniers un antidictionnaire permettant d'éliminer les « passe-partout » comme *important, grand, nécessaire*, et de garder ceux qui pourraient être nominalisés avec précision et profit, comme *irakien, hydraulique, antisémite*. Un antidictionnaire des prénoms a également été appliqué.

L'environnement NeuroNav offre des fonctions utiles au contrôle « transversal » rapide d'un vocabulaire de plusieurs dizaines de milliers de termes : l'expansion lexicale (cf. Fig. 3) d'une chaîne de caractères permet de repérer des variantes et fautes d'orthographe récurrents qu'il pourrait être intéressant de fusionner, l'expansion sémantique (termes les plus proches au sens du cosinus dans l'espace distributionnel phrases-termes) permet de repérer des synonymes lexicalement éloignés, par exemple ici la liste des interviewers comme *Anne Sinclair, Yves Mourousi*, … que nous avons mis de côté et fusionnés sous le terme *interviewer*. Ces fonctions permettent un nettoyage semi-automatique du vocabulaire dans un temps raisonnable, en s'intéressant en premier lieu aux mots et expressions les plus récurrentes : quelques demi-journées de travail dans notre cas pour éliminer les principaux termes à fonction rhétoriques, éliminer les principaux termes contenant des nombres (en dehors d'une liste contenant *14_juillet, 16eme_parallele, 21eme_siecle*…) et autoriser une extraction de 200 thèmes centrés uniquement sur le contenu sémantique traité, en deux ou trois va-et-vient entre nettoyage du vocabulaire et clustering, comme décrit plus haut pour l'extraction des thèmes rhétoriques. Cette procédure « informée par le sens », semi-

automatique mais rapide, permet à notre avis une réduction de dimension de meilleure qualité qu'avec une procédure aveugle de type Latent Semantic Analysis.

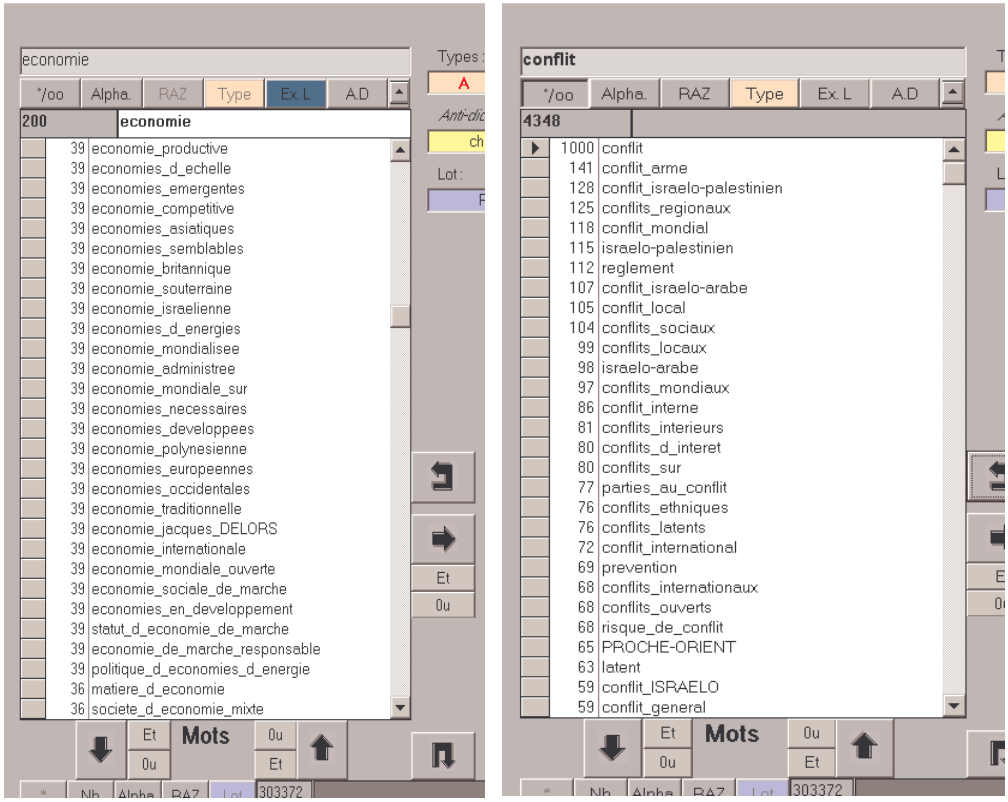

Fig. 3 – Expansion lexicale de* économie *et sémantique de* conflit *sous NeuroNav.

Un indicateur simple et imprévu de débuts de paragraphe :

Le nombre de mots significatifs issu de notre nettoyage de vocabulaire s'est révélé fortement lié au problème posé : parmi les phrases dépourvues de mots sémantiquement significatifs, pratiquement aucune ne s'est trouvée en début de paragraphe.

En termes de résultats quantitatifs, l'optimum a été trouvé pour la valeur de coupe 5 (0 pour le critère flou, équivalent à une non prise en compte du critère[7]) :

	Fstrict	Fflou1
Appr.	0,154	0,316
Ajust.	**0,157**	**0,318**
Test	0,142	0,296

[7] Ceci équivaut à la stratégie « paresseuse-gagnante » consistant à étiqueter positifs tous les éléments de l'ensemble de test, si l'on sait qu'il y a moins de positifs que de négatifs…

Extraction de 200 thèmes sémantiques

Un compromis entre réduction de l'espace des données et minimisation des phrases « laissées pour compte » (pour cause de valeurs de projection nulles sur tous les axes) a consisté à choisir 200 comme nombre de thèmes sémantiques à extraire, à partir de 80% des paragraphes de l'ensemble d'apprentissage (de façon à laisser quelques milliers de phrases pour les procédures ultérieures d'ajustement et de test d'apprentissage). Tous étaient clairement interprétables (cf. Tab. 1). Les phrases ont ensuite été projetées sur les 200 axes obliques de thèmes, passivement, en éléments supplémentaires (fonction *lot secondaire* de NeuroNav).

0.517	174	**magistrat**
0.465	1399	justice
0.403	38	cour_de_cassation
0.397	3	cassation
0.353	486	juge
0.343	21	presomption_d_innocence
0.327	6	innocence
0.325	15	procureur
0.305	81	sceau
0.302	26	parquet
0.285	13	presomption
0.264	25	procureur_general
0.264	39	penal
0.257	60	conseil_superieur
0.256	70	magistrature

0.498	737	**science**
0.473	1783	recherche
0.447	581	scientifique
0.369	320	chercheur
0.291	47	recherche_fondamentale
0.258	1128	connaissance
0.252	247	medecine
0.249	24	recherche_appliquee
0.224	12	reflexion_ethique
0.221	133	biologie
0.214	87	recherche_scientifique
0.197	4	biomedical
0.196	197	decouverte
0.190	248	institut
0.190	241	savant

TAB 1 – Deux exemples d'illustration de thèmes sémantiques par leurs listes de termes. En colonnes, la projection du terme sur l'axe du thème et sa fréquence totale .

Changement de niveau : passage aux phrases.

Nous avons envisagé deux possibilités pour cette transition :

1) Passage au niveau phrases *avant* analyse :

Dans ce cas, on n'utiliserait que l'information « nombre moyen de phrases par paragraphe (=15 sur tout le corpus, mais il serait possible de spécifier plus finement selon l'auteur) pour constituer 200 000 fenêtres glissantes d'empan 15 et de pas 1 phrase, chacune de ces nouvelles unités textuelles regroupant l'ensemble des termes des 15 phrases. Nous avions déjà utilisé cette méthode avec succès dans l'analyse non supervisée du discours « au kilomètre » d'un paranoïaque (Schepens et al. 2004). Mais, outre le respect dû à la parole présidentielle, outre le coût en temps de calcul sur 200 000 unités, cette solution nous a semblé éloignée de l'esprit de notre réponse DEFT, qui était idéalement de tenter d'approcher les catégories thématiques établies par les documentalistes pour ce corpus, en tant qu'étape intermédiaire. C'est pourquoi nous avons préféré – à tort ou à raison, seule l'expérience pourrait trancher – la méthode ci-après.

2) Passage au niveau phrases *après* analyse :

En préambule, notons que les résultats issus de notre méthode des K-means axiales relèvent à la fois du domaine symbolique et du domaine numérique. Pour chaque texte, on obtient en effet :

- un ensemble de valeurs de projection (des cosinus) sur les K axes de thèmes – axes non orthogonaux, contrairement à la plupart des méthodes d'analyse factorielle et autres Latent Semantic Analysis. La projection la plus grande détermine l'étiquette (le thème dominant).
- une étiquette qualitative (une parmi K) d'appartenance à un thème, parmi K thèmes extraits.

On sépare ainsi l'information pertinente du bruit, avec une bonne précision pour les valeurs de projections élevées (mais un rappel difficile à la fois à définir et à appréhender en l'état de l'art actuel). À preuve les 200 thèmes que nous avons extraits à partir de 80% des paragraphes du corpus d'apprentissage, tous remarquablement clairs d'interprétation (cf. exemples plus haut). Notre idée de « déconstruire » cette information en calculant les projections de toutes les phrases sur les 200 axes semble malheureusement avoir reconstitué le bruit dont on s'était départi au niveau supérieur.

Nous avons utilisé sur ces phrases des fenêtres glissantes telles que définies plus haut (section 4.1, hypothèse H2), avec leurs paramètres à la fois fixes sur l'ensemble du corpus et contrôlés par l'optimisation du Fscore. À défaut d'éléments théoriques utilisables ici concernant l'utilisation mixte d'éléments symboliques et numériques (Kodratoff et Diday 1991), nous avons défini pour les similarités entre phrases un « Lego » d'éléments divers (nombre fixe - ou au dessus d'un certain seuil - de thèmes dominants et/ou moyens, taille et poids de listes de thèmes, méthodes et résultats de comparaison de listes) dont l'assemblage a été optimisé lui aussi par le Fscore. Sur cette base, c'est l'utilisation des « différences de différences » de similarité entre phrases qui s'est avérée la plus efficace. Cette cascade empirique a permis de parvenir à un Fscore strict de 0.163 ±.002 sur nos trois ensembles, médiocre mais nettement supérieur au score de 0.067 si l'on avait répondu au hasard, ou au score de 0.125 pour la stratégie « paresseuse-gagnante » d'étiquetage positif généralisé.

Une explication pour ces modestes performances au regard des efforts fournis, certainement pas la seule, pourrait tenir à la forte variabilité du nombre de phrases par paragraphe (hétérogénéité du passage entre niveaux, encore plus marquée dans le corpus *Ouvrage scientifique* de ce défi). Compte tenu des liens constatés avec la variable « auteur présumé » (cf. figure 1) un pis-aller aurait été de faire intervenir ce nombre moyen pour chaque auteur. D'autres solutions (fenêtres glissantes à balayage variable « ver de terre » ?) pourraient aussi être explorées.

5 - Apprentissage final sur les 5 indicateurs synthétisant nos points de vues

Nos 5 indicateurs forment un nouveau tableau de données de 303 000 lignes (pourvues d'étiquettes *phrase de début de paragraphe / autre*) et 5 colonnes, dont l'une (col. 5, indicateur de rupture de thème sémantique) est quasi-orthogonale aux autres, alors que les colonnes 1 et 2 (débuts de phrases) sont à la fois très liées entre elles (41°) et modérément liées à la col. 3 (thèmes rhétoriques, 71° et 73°). Cette dernière n'est pas dépourvue de relation avec la col. 4 (nombre de mots significatifs, 80°). Pris ensemble, ces indicateurs balisent bien nos trois points de vues complémentaires sur le corpus mentionnés en section 2. Ce sont tous des fonctions croissantes de la probabilité d'être en début de paragraphe. On a désormais affaire à un tableau « idéal » pour méthodes d'apprentissage, avec beaucoup de lignes et peu de colonnes.

Nous avons utilisé la méthode des règles de décision (Zighed et Rakotomalala 2000) pour nos trois réponses au défi. Le parcours de grilles de seuils sur ces indicateurs permet d'optimiser le Fscore strict ou flou, au choix, suivant les principes exposés plus haut.

. *Réponse N° 3 :*

Comme pratiqué généralement, nous avons cherché les meilleurs seuils pour une règle de type « SI col1≥seuil1 ET col2≥seuil2 ET etc. ALORS phrase de début », en optimisant Fstrict, ce qui nous a donné les valeurs :

Appr. 0.404
Ajust. 0.284
Test 0.374
Test DEFT 0.284

Ce type de règle ne comportant que des ET définit un hyperrectangle frontière entre points positifs et négatifs dans la portion de l'espace opposée à l'origine. De façon symétrique, une règle ne comportant que des OU définit un hyperrectangle frontière entre points positifs et négatifs dans la portion de l'espace contenant l'origine. Nous avons essayé cette variante, et constaté que les résultats n'étaient que légèrement inférieurs :

Appr. 0.375
Ajust. 0.297
Test 0.355

D'où l'idée que la vérité pourrait se situer dans des frontières moins simplistes, combinant les critères ET et OU. La combinaison testée dans nos réponses 2 et 1 est issue du bon sens et de notre connaissance de la signification des indicateurs – elle pourrait aussi faire l'objet d'une optimisation automatique.

. *Réponse N° 2 :*

Règle : SI nb. mots significatifs ≥ seuil1
ET ($début_1$ phrase ≥ seuil2 OU $début_2$ phrase ≥ seuil3 OU rhéto ≥ seuil4)
ET rupture sémantique ≥ seuil5
ALORS phrase de début

L'optimisation de Fstrict a donné :

Appr. 0.395
Ajust. 0.301
Test 0.379
Test DEFT 0.282

. *Réponse N° 1 :*

Avec la même règle l'optimisation de Fflou1 a donné les valeurs :

Appr. 0.412
Ajust. 0.335
Test 0.345
Test DEFT 0.342

L'inconvénient de cette méthode de règles de décision est qu'elle taille « à la serpe » des frontières nécessairement parallèles aux axes, et peut très bien exclure un indicateur un peu plus faible que les autres et non totalement orthogonal en plaçant son seuil à la borne

inférieure. C'est pourquoi nous aurions préféré des méthodes traçant des frontières plus nuancées, tirant le meilleur de *l'ensemble* des indicateurs.

Des essais avec la méthode des K plus proches voisins (meilleur K : 3) n'ont pas donné de résultats supérieurs.

Par contre, nous avons fait après-coup des essais avec validation croisée simple et optimisation des paramètres en utilisant la méthode SVM au moyen de la librairie *libsvm* (Chang et Lin 2001) ; avec le choix d'un noyau RBF de paramètre gamma=1/2, du paramètre de compromis marge / erreurs admissibles c = 0.004, du poids des exemples positifs = 14, nous avons obtenu sur divers échantillons tirés au 1/25 ou 1/10 des Fscore de test entre 0.38 et 0.41. Nous avons confronté ces résultats a priori inespérés au véritable ensemble de test, mais n'avons obtenu qu'une amélioration de 0.03 environ sur les Fscore strict et flous…

6 Conclusion générale

Notre approche d'apprentissage synthétique final à partir de plusieurs points de vues plus ou moins orthogonaux a été validée, de façon un peu sous-estimée dans nos trois réponses DEFT si on en croit le dernier test mentionné ci-dessus.

L'utilisation de nos techniques non-supervisées (indexation contrôlée NeuroNav, méthode de clustering K-means axiales) a été validée pour la construction d'indicateurs de « thèmes rhétoriques » et d'« intensité sémantique » (nombre de mots significatifs).

Pour ce qui concerne la détection des ruptures thématiques, si le problème de la transition entre les niveaux phrases et paragraphes a bien été identifié, les solutions trouvées, sans être déshonorantes, demandent encore beaucoup de recherche et de réflexion. Il s'agit là d'un point fondamental pour l'apprentissage de données *complexes*.

A plus court terme, nos résultats auraient pu être améliorés :

- par la prise en compte de l'étiquette « fin de paragraphe »,
- par la prise en compte des auteurs principaux dans la construction de nos fenêtres glissantes sémantiques (éventuellement par l'induction d'éléments de chronologie pour le sous-corpus Chirac),
- par une validation par fraction tournante au sein de chaque bloc de même auteur principal,
- par la prise en compte de la séquentialité des phrases pour renforcer la fiabilité des indicateurs,
- par l'utilisation de méthodes plus élaborées que les règles de décision dans l'apprentissage synthétique final.

Ces améliorations partielles pourraient peut-être porter le Fscore jusqu'à 0.5-0.6. Mais pourra-t-on sans progrès décisifs en matière d'intégration multiniveau franchir le saut jusqu'aux valeurs supérieures à 0.9, au moins pour le F flou d'ordre 1, exigées au minimum par toute application opérationnelle ?

Concernant la possibilité de généraliser notre démarche à d'autres corpus textuels, voire à d'autres problèmes d'apprentissage que le texte, nous pensons que cela est possible dès lors que les objets à identifier sont nombreux et descriptibles selon plus d'un point de vue. Par exemple, des articles scientifiques peuvent être décrits par leurs citations, en plus de leurs multiples aspects textuels. Le corpus Euro-Lex du défi DEFT'06 présente à nos yeux moins de difficultés de transition entre niveaux que celui des discours présidentiels, puisque le

passage entre phrase et article de loi y semble trivial ; faute de temps pour nous y confronter, notre impression est que les aspects formels et rhétoriques, sans doute très redondants, y prédominent pour établir le découpage en lois, rendant de ce fait notre approche moins productive par rapport à un apprentissage plus unidimensionnel. A l'opposé, c'est la faible quantité de textes disponibles en regard de leur extrême variété rhétorique et sémantique, ainsi que la complexité de leur structuration en chapitres, sections, sous-sections, etc. qui a contribué à nous faire abandonner l'application de notre méthode à la tâche de découpage de l'ouvrage scientifique proposée dans DEFT'06.

Références

Chih-Chung Chang and Chih-Jen Lin (2001) LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Cornuéjols A. et Miclet M. (2002) Apprentissage artificiel : concepts et algorithmes, Eyrolles, 2002.

DEFT'2006, Défi Fouille de Textes, 2006. Cf. http://www.lri.fr/ia/fdt/DEFT06/

Dussart C. et Petit C. (2005) Méta-apprentissage d'expériences : nouvelles voies en Data Mining, Vuibert, 2005.

Kodratoff Y. et Diday E. (1991) Induction Symbolique et Numérique à partir de données, Cépaduès éditions, 1991.

Lelu A. (1994), Clusters *and* factors: neural algorithms for a novel representation of huge and highly multidimensional data sets, in E. Diday, Y. Lechevallier & al. eds., New Approaches in Classification and Data Analysis, p. 241-248, Springer-Verlag, Berlin, 1994

Lelu A. et Aubin S. (2001), Vers un environnement complet de synthèse statistique de contenus textuels, séminaire ADEST, 2001. http://web.upmf-grenoble.fr/adest/seminaires/lelu02/ADEST2001_SA_AL.htm

Schepens P., Lelu A., Viprey J.-M. (2004), Essais d'analyse de discours sur des corpus d'entretiens cliniques, rapport interne LASELDI, 50p., Université de Franche-Comté, 2004.

Ziadé T. (2006), Programmation Python : syntaxe, conception et optimisation, Eyrolles, 2006.

Zighed D.A., Rakotomalala R. (2000) Graphes d'induction, Apprentissage et Data Mining, Hermès Sciences Publications, 2000.

Summary

Our answer to the challenge of detecting the borders of thematic textual segments in French presidents' speeches lays upon the building of a small number of numerical transition indicators, as orthogonal as possible, each embedding a unique point of view: formal metadata, rhetorical, semantic, … Three of them were issued from a dramatic dimensionality reduction process, from tens thousands words to one or two hundred significant and interpretable dimensions (oblique positive factors), in an unsupervised way. The semantic indicator was critical in that it involved a 2-level disaggregation/ aggregation problem, for which we designed a solution, though unsatisfactorily. The final higher level learning phase of our 5 indicators relied on decision rules, and resulted in good strict and fuzzy scores, compared to the other challengers.