Nouvelles approches statistiques pour la classification dynamique des textes et l'analyse des changements et de la créativité linguistique

Jean-Charles Lamirel
Equipe Synalp – Laboratoire LORIA, Nancy, France
Email : lamirel@loria.fr

Résumé

Face à l'évolution de la notion de texte et à la croissance continuelle de l'information textuelle, de multiple nature, disponible en ligne, et à l'évolution des langues proprement dites, un des enjeux important pour les linguistes, pour pouvoir étayer ou construire des hypothèses et valider des modèles liés aux changements et à la créativité linguistiques, est celui de pouvoir disposer d'outils d'analyse textuelle efficaces, capables de s'adapter à de gros volumes de données, souvent de nature hétérogène, et distribuée. Nous proposons dans cette communication de nous intéresser à de nouvelles méthodes statistiques qui s'inscrivent dans ce cadre.

Les mesures statistiques les plus utilisées en analyse textuelle sont des mesures distributionnelles, principalement basées sur l'entropie, ou sur la métrique du Chi2. Ces mesures sont de plus généralement exploitées dans le cadre d'une analyse globale. Il est cependant couramment admis que ce type d'approche présente de fortes limitations pour la détection du changement, où l'effectif des informations nouvelles est fortement déséquilibré par rapport à celui des informations courantes. Nous avons récemment proposé une approche alternative basée sur la métrique de maximisation des traits et sur l'analyse multi-vues qui ne présente pas ces inconvénients. Nous avons montré qu'elle permettait de traiter très efficacement de nouveaux problèmes liés à l'analyse des données textuelles, comme le suivi de sujets et la détection de nouveauté dans des données changeantes au cours du temps, tout en s'adaptant à l'analyse discriminante traditionnelle, fortement exploitée en linguistique, et corrélativement, à la catégorisation des textes, à la stylométrie, et, à l'assistance à la génération de lexiques, avec des performances très supérieures aux méthodes classiques susmentionnées. Un autre de ses avantages déterminants est qu'elle permet de travailler en se basant sur des approches sans référentiel (non supervisées), aussi bien qu'avec des approches classificatoires traditionnelles. Au final, elle affranchit également l'analyste de l'exploitation de paramètres.

Nous présentons les principes généraux de la méthode et nous revenons, à l'aide de plusieurs exemples, sur ses différents domaines d'application dans le cadre supervisé. De manière additionnelle, nous montrons comment cette méthode cette méthode permet d'identifier et de visualiser de manière pertinente la pluridisciplinarité et les champs pluridisciplinaires dans les corpus.

1. Introduction

La catégorisation automatique de textes (CAT) vise à regrouper, souvent selon des thèmes communs, les documents ayant des caractéristiques spécifiques et homogènes (Cohen et Hersh, 2005). La première étape de ce type de catégorisation est la transformation des documents en une représentation appropriée pour le classifieur. Cette transformation vise à pondérer et à réduire l'espace de représentation des documents tout en ménageant la possibilité de discriminer entre ces derniers. Elle comprend usuellement des opérations de suppression des mots vides, de lemmatisation, de sélection et de pondération des descripteurs. La deuxième étape est l'apprentissage : le système apprend à classer les documents selon un modèle de classement où les classes sont prédéterminées et les exemples sont connus et correctement étiquetés d'avance.

La catégorisation automatique de textes a été l'un des domaines les plus étudiés en apprentissage automatique (Hillard et al., 2007). En conséquence, une grande variété d'algorithmes de classification ont été développés et/ou évalués, souvent dans des applications telles que le filtrage des mails (Cormack, 2007) ou l'analyse des opinions et des sentiments (Pang et Lee, 2008). Dans le domaine des sciences sociales, l'apprentissage automatique a été utilisé dans la classification d'actualités (Purpura et Hillard, 2006, Evans et al., 2007), ou des blogues (Durant et Smith, 2007). Parmi les méthodes d'apprentissage les plus souvent utilisées, figurent les réseaux de neurones (Wiener et al., 1995, Schütze et al., 1995), les K-plus-proches-voisins (K-PPV) (Yang et Chute, 1994), les arbres de décision (Lewis et Ringuette, 1994, Apte et al., 1998), les réseaux bayésiens (Lewis, 1992, Joachims, 1997), les machines à vecteurs supports (SVM) (Joachims, 1998), et plus récemment, les méthodes basées sur le boosting (Schapire, 1998, Iyer et al., 2000). Bien que beaucoup de méthodes développées dans le domaine de la catégorisation automatique de textes aient permis d'atteindre des niveaux de précision appréciables lorsqu'il s'agit de textes à structure simple (par ex. courriels, résumés, etc.), il reste néanmoins encore défis à relever dans les cas les plus complexes ou les description des documents sont bruitées et/ou basées sur un nombre important de descripteurs, les classes s'avèrent relativement similaires, et la répartition des exemples entre les différentes classes d'apprentissage n'est pas équilibrée.

Nous montrons ci-après que l'exploitation d'une méthode de sélection de variables basée sur notre métrique de maximisation des traits permet à la fois de gérer le déséquilibre des classes, la similarité et de réduire la complexité de la tâche de classification et qu'elle permet de mettre en évidence des profils de classes, ce qui s'avère particulièrement utile dans le cadre de la classification des textes en général, et dans celui de la stylométrie, en particulier. A la section 2, nous présentons le contexte d'exploitation des méthodes de sélection, ainsi que les classes usuelles de méthodes, puis nous présentons, à la section 3, la méthode alternative basée sur les traits que nous proposons en illustrant son fonctionnement à l'aide d'un exemple simple. A la section 4, nous présentons les corpus textuels de référence sur lesquels nous menons nos expérimentations. Les résultats

obtenus sont décrits à la section 5. Nos conclusions et perspectives sont énoncées dans la dernière section.

2. La sélection de variables

Depuis les années 1990, les progrès de l'informatique et des capacités de stockage permettent la manipulation de très gros volumes de données: il n'est pas rare d'avoir des espaces de description de plusieurs milliers, voire de dizaines de milliers de variables. On pourrait penser que les algorithmes de classification sont plus efficaces avec un grand nombre de variables. Toutefois, la situation n'est pas aussi simple que cela. Le premier problème qui se pose est l'augmentation du temps de calcul. En outre, le fait qu'un nombre important de variables soient redondantes ou non pertinentes pour la tâche de classification perturbe considérablement le fonctionnement des classifieurs. De plus, la plupart des algorithmes d'apprentissage exploitent des probabilités et les distributions de probabilités peuvent alors être difficiles à estimer dans le cas de la présence d'un très grand nombre de variables. L'intégration d'un processus de sélection de variables dans le cadre de la classification des données de grande dimension devient donc un enjeu central. Ceci est d'autant plus vrai qu'il est également nécessaire, pour des raisons de synthèse, de mettre en avant les variables privilégiées lors de la visualisation des résultats de classification.

Dans la littérature, trois types d'approches pour la sélection de variables sont principalement proposés: les approches directement intégrées aux méthodes de classification, dites « embedded », les méthodes basées sur des techniques d'optimisation, dites « wrapper », et finalement, les approches de filtrage, telle que la méthode du Chi2. Des états de l'art exhaustifs des différentes techniques ont été réalisés par de nombreux auteurs, comme Ladha et al. (Ladha et Deepa, 2011), Bolón-Canedo et al. (Bolón-Canedo et al., 2012), Guyon et al. (Guyon et Elisseeff, 2003) ou Daviet (Daviet, 2009). Pour avoir un aperçu de ces méthodes, le lecteur se réfèrera aux articles mentionnés, ainsi qu'à (Lamirel et al. 2013).

3. Maximisation des traits pour la sélection de variables

3.1. Principe de la métrique de maximisation des traits en apprentissage non supervisé

La maximisation des traits (F-max) est une métrique non biaisée d'estimation de la qualité d'une classification non supervisée (clustering) qui exploite les propriétés des données associées à chaque cluster sans examen préalable des profils de clusters. Cette métrique a été initialement proposée dans (Lamirel et al. 2004). Son principal avantage est d'être tout à fait indépendante des méthodes de classification et de leur mode de fonctionnement.

Considérons un ensemble de données D représenté par un ensemble de variables F, et un ensemble de clusters C résultant d'une méthode de clustering. La métrique de maximisation des traits favorise les clusters avec une valeur maximale de F-mesure de trait.

La F-mesure de trait d'une variable *f* associée à un cluster *c* est définie comme la moyenne harmonique du rappel de trait et de la prépondérance de trait elle mêmes définies comme suit :

$$FR_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c' \in C} \sum_{d \in c'} W_d^f} \quad (1)$$

$$FP_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{f' \in F_c, d \in c} W_d^{f'}} \quad (2)$$

où W_d^f représente le poids de la variable *f* pour les données *D* et F_c représente l'ensemble des caractéristiques des données associées au cluster *c*.

Deux applications importantes de la métrique de maximisation des traits sont liées à l'estimation de la qualité du clustering et au clustering incrémental. Plus de détails sur ces applications sont donnés dans les références ((Lamirel et al., 2011, Falk et al., 2012). Des expériences préalables menées sur l'étiquetage de clusters avaient également montré que la métrique présentait des capacités de discrimination équivalentes à celle du Khi 2, tout en ayant des capacités de généralisation très sensiblement supérieures (Lamirel et al., 2008).

3.2. Adaptation de la métrique de maximisation des traits pour la sélection de variables en apprentissage supervisé

Tenant compte de la définition de base de la métrique de maximisation des traits présentée dans la section précédente, son exploitation pour la tâche de sélection de variables dans le contexte de l'apprentissage supervisé devient un processus simple, dès lors que cette métrique générique peut s'appliquer sur des données associées à une classe aussi bien qu'à celles qui sont associées à un cluster. Le processus de sélection peut donc être défini comme un processus non paramétré basé sur les classes dans lesquelles une variable de classe est caractérisée en utilisant à la fois sa capacité à discriminer une classe donnée (index $FR_c(f)$) et de sa capacité à représenter fidèlement les données de la classe (index $FP_c(f)$).

L'ensemble S_c des variables qui sont caractéristiques d'une classe donnée *c* appartenant à un ensemble de classes *C* se traduit par:

$$S_c = \{ f \in F_c \mid FF_c(f) > \overline{FF}(f) \; and \; FF_c(f) > \overline{FF}_D \} \quad (3)$$

où $\overline{FF}(f) = \sum_{c' \in C} FF_{c'}(f) / \left| C_{/f} \right|$ et $\overline{FF}_D = \sum_{f \in F} \overline{FF}(f) / |F|$.

et $C_{/f}$ représente un sous-ensemble de *C* aux classes dans lesquelles la variable *f* est représentée.

Enfin, l'ensemble S_C de toutes les variables sélectionnées est le sous-ensemble de F défini comme:

$$S_C = \bigcup_{c \in C} S_c \quad (4)$$

Les variables qui sont jugées pertinentes pour une classe donnée sont les variables dont les représentations sont meilleures que leurs représentations moyennes dans toutes les classes, et meilleures que la représentation moyenne de toutes les variables, en termes de F-mesure de trait.

Dans le cadre spécifique du processus de maximisation des traits, une étape d'amélioration par contraste peut être exploitée en complément de la première étape de sélection. Le rôle de cette étape est d'adapter la description de chaque donnée aux caractéristiques spécifiques de leurs classes associées qui ont été précédemment mises en évidence par l'étape de sélection (Lamirel et al. 2013). Dans le cas de notre métrique, cela consiste à modifier le schéma de pondération des données pour chaque classe en prenant en considération le gain d'information fourni par la F-mesure de trait des variables, localement à cette classe.

Le gain d'information est proportionnel au rapport entre la valeur de la F-mesure d'une variable dans la classe et la valeur moyenne de la F-mesure de variable dans toute la partition. Pour une variable f appartenant à l'ensemble S_c des variables sélectionnées pour une classe *c*, le gain $G_c(f)$ susmentionné peut s'exprimer comme:

$$G_c(f) = (FF_c(f)/\overline{FF}(f))^k \quad (5)$$

où *k* est un facteur de magnification qui peut être optimisé en se basant sur la pertinence obtenue.

Les *variables actives* d'une classe sont celles pour lesquelles le gain d'information est supérieur à *1* dans celles-ci. Etant donné que la méthode proposée est une méthode de sélection et de contraste basée sur les classes, le nombre moyen de variables actives par classe est donc comparable au nombre total de variables sélectionnées dans le cas des méthodes de sélection usuelles.

Nous donnons ci-après un exemple de fonctionnement de la méthode sur une base d'un exemple-jouet comprenant deux classes (*Homme (M), Femme (F)*) décrites par trois variables (*Taille_Nez (N), Longueur_Cheveux (C), Taille_Pieds (P)*). Le tableau 1 présente les données source et montre comment s'opère le calcul de la F-mesure de trait de la variable *Taille_Pieds* vis-à-vis de la classe *Homme*.

Comme le montre le tableau 2, la seconde étape du processus consiste à calculer les valeurs de moyenne marginale de F-mesure de trait pour chaque variable et la moyenne globale de F-mesure, toutes variables et toutes classes confondues. Les variables dont la

F-mesure est systématiquement inférieure à la moyenne globale sont éliminées. La variable *Taille_Nez* est ainsi supprimée.

Taille Pieds	Longueur Cheveux	Taille Nez	Classe
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	F
6	25	6	F
5	25	5	F

$FR(P,M) = 27/43 = 0.62$

$FP(P,M) = 27/78 = 0.35$

$FF(P,M) = \frac{2(FR(P,M)\times FP(P,M))}{FR(P,M)+FP(P,M)} = 0.48$

Tab 1. : Données-exemple et calcul de la F-mesure de trait.

	F(x,M)	F(x,F)	$\overline{F(x,.)}$
Longueur Cheveux	0.39	0.66	0.53
Taille Pieds	0.48	0.22	0.35
Taille Nez	0,3	0,24	0,27

$\overline{F(.,.)}$
0.38

Tab 2. : Tableau des F-mesures de trait de variables et de leurs moyennes.

Les variables restantes (sélectionnées) sont considérées actives dans les classes dans lesquelles la F-mesure de trait est supérieure à la moyenne marginale :

1. *Taille_Pieds* est active dans la classe *Homme*,
2. *Longueur_Cheveux* est active dans la classe *Femme*.

Le facteur de contraste met en évidence le degré d'activité/passivité des variables sélectionnées par rapport à leur F-mesure moyenne marginale dans les différentes classes. Le tableau 3 montre comment le contraste est calculé sur l'exemple présenté. Dans le contexte de cet exemple, le contraste pourra ainsi être considéré comme une fonction qui aura virtuellement les effets suivants :

1. Augmenter la longueur des cheveux des femmes,
2. Augmenter la taille des pieds des hommes,
3. Diminuer la longueur des cheveux des hommes,
4. Diminuer la taille des pieds des femmes.

La dernière étape du processus consiste à appliquer le contraste sur les données sources. Le tableau 4 montre qu'une fois les données contrastées la séparation entre les classes *Homme* et *Femme* devient possible, alors qu'elle ne l'était pas sur les données originales. Ce processus s'apparente à un processus de transformation non linéaire sur les données.

	F(x,M)	F(x,F)	$\overline{F(x,.)}$
Long. Cheveux	0.39	0.66	0.53
Taille Pieds	0.48	0.22	0.35

	C(x,M)	C(x,F)
Long. Cheveux	0.39/0.53	0.66/0.53
Taille Pieds	0.48/0.35	0.22/0.35

	C(x,M)	C(x,F)
Long. Cheveux	0.74	1.25
Taille Pieds	1.37	0.63

Tab 3. : Principe de calcul du contraste sur les variables sélectionnées et résultats obtenus.

Taille Pieds	Longueur Cheveux	Classe
9	5	M
9	10	M
9	20	M
5	15	F
6	25	F
5	25	F

Données originales (variables retenues)

Taille Pieds	Longueur Cheveux	Classe
12,33	3.7	M
12,33	7.4	M
12,33	14.8	M
3.15	18.75	F
3,78	31.25	F
3.15	31.25	F

Données contrastées

Tab 4. : Données réduites aux variables retenues avant et après contraste.

Un exemple concret de l'intérêt de l'exploitation de la sélection de variables basée sur la maximisation des traits est donné par la tâche d'assistance à la validation des brevets du projet QUAERO. Cette tâche consistait à générer un aide aux experts dans leur tâche d'évaluation de la nouveauté d'un brevet fondée sur l'assignation automatique des papiers scientifiques pertinents liés avec les codes de la de classification des brevets. Dès lors que l'apprentissage était basé sur les citations extraites des brevets qui sont habituellement associées à une hiérarchie des codes de classification ayant différents niveaux de généralité, en premier lieu, il n'y avait aucune garantie d'une répartition homogène des citations (c.-à-d. les échantillons d'apprentissage) parmi les codes, en second lieu, il y avait de fortes chances d'avoir des citations similaires dans différentes classes. Cette tâche

soulevait donc de nouveaux défis dans le domaine de la classification, en particulier celui de devoir traiter des données très déséquilibrées appartenant à des classes fortement similaires entre elles (Hajlaoui et al. 2012). Elle n'a pu fournir des résultats satisfaisants et exploitables par les experts (une faible confusion entre les classes s'avérait naturellement indispensable dans ce cadre) qu'après l'exploitation de mécanismes de sélection de variable basés sur la maximisation des traits. En effet, dans ce contexte, cette méthode a permis d'améliorer les performances de la classification de plus de 90%, alors que toute les méthodes concurrentes, notamment la méthode du Chi2, se sont révélées totalement inopérantes, voir néfastes aux performances (Lamirel et al. 2013).

4. Les données expérimentales

4.1. Corpus Deft'05

Une étude des discours de Mitterrand a été menée en 2000 par Habert et al., en comparaison avec les discours de De Gaulle à partir d'un corpus d'interventions radio-télévisées. L'étude des spécificités de la distribution des traits linguistiques fait apparaître les traits dominants suivants chez Mitterrand :

adverbe négatif, pronom personnel première personne singulier, indicatif présent, article indéfini, passé composé, tiret, pronom démonstratif, c'est (à l'indicatif présent), deux points, pronom personnel, nombre cardinal, point d'interrogation, il y a (à l'indicatif présent), il faut (à l'indicatif présent), pronom personnel on, pouvoir à l'indicatif présent.

Nous avons utilisé un corpus de discours des présidents Chirac et Mitterrand issu du défi DEFT'05[1]. Ce défi était basé sur un corpus d'extraits d'allocutions de F. Mitterrand introduites dans des textes de J. Chirac, et comportait les trois tâches suivantes :

1. Identifier les phrases de Mitterrand dans un corpus ne comportant ni années, ni noms de personnes,
2. Identifier les phrases de Mitterrand dans un corpus ne comportant pas d'années,
3. Identifier les phrases de Mitterrand dans un corpus avec les années et les noms de personnes.

Le corpus utilisé comporte 73255 phrases de J. Chirac et 12320 phrases de F. Mitterand.

La présentation des corpus ainsi que la synthèse des résultats du défi Deft'05 a été publié lors de la conférence TALN'05 (Alphonse et al. 2005).

Les meilleurs résultats (tableau 5) ont été obtenus par une équipe du LIA d'Avignon (El-Bèze et al. 2005), en utilisant des méthodes probabilistes (chaines de Markov, et modèles bayésiens). Les auteurs ne lemmatisent pas et ne filtrent pas les données afin d'éviter un processus de prétraitement qu'ils estiment lourd pour certaines langues (même s'ils reconnaissent que cela dégrade légèrement les résultats F-score de 0.84 au lieu de 0.88).

[1] https://www.lri.fr/~aze/fdt/DEFT05/

Rigouste et al. (2005) utilisent également un algorithme de Viterbi (modèle de Markov caché). Ils mettent en œuvre un modèle non supervisé pour une tâche de fouille de textes supervisée : à partir du corpus de test ils identifient des thématiques (et leur distribution) dans les discours de Chirac et de Mitterrand. Chaque document est représenté par un sac de mots, et l'idée majeure est qu'il est plus aisé de détecter des ruptures thématiques connaissant les sujets abordés par les locuteurs. Les discours de chaque président sont modélisés par plusieurs thèmes.

L'équipe du LORIA (Pierron et al. 2005) utilise un classifieur bayésien dont les réponses sont pondérées et seuillées. Ils créent un indice qui est égal à 2*probabilité (phrase soit de Mitterrand)+(1-probabilité(phrase soit de Chirac)). Le seuil est ajusté pour optimiser la valeur de F-score sur l'ensemble de test.

Labadié et al. (2005) combinent une adaptation de C99 (C99 est une méthode de segmentation thématique non-supervisée qui s'appuie sur des calculs de similarités entre phrases (Choi, 2000)) et une classification bayésienne.

Equipe	Tâche 1	Tâche2	Tâche 3
Elbeze_LIA	0.87	0.88	0.88
Rigouste_ENST	0.86	0.85	0.87
Pierron_LORIA	0.82	0.82	0.82
Labadie_LIA	0.76	0.74	0.75
Maisonnasse_CLIPS	0.75	0.75	0.76
Kerloch_LIP6	0.73	0.79	0.79
Hernandez_LIMSI	0.56	0.56	0.57
Plantie_LGI2P	0.49	0.52	0.51
Hurault-Plantet_LIMSI	0.49	0.56	0.56
Hauche_LIRMM	0.32	0.31	0.31
Moot_LABRI	0.18	0.18	0.42

Tab 5. : Résultats des meilleures exécutions en termes de F-mesure.

En 2006, Michèle Jardino (Jardino 2006) essaye d'identifier les auteurs à partir de n-grammes de caractères ou de mots. Les meilleurs résultats sont obtenus pour des caractères avec n=4 (F=0.75) alors qu'avec les mots elle obtient pour n=3 à 6 F=0.68.

4.2. Corpus UCI Amazon

Le Corpus Amazon™ (AMZ) est un ensemble de données UCI (Bache et Lichman, 2013) dérivé des avis de clients du site web Amazon et exploitable pour l'identification des auteurs. Pour examiner la robustesse des algorithmes de classification avec un grand nombre de classes-cibles, 50 des utilisateurs les plus actifs en termes de commentaires postés dans ces newsgroups sont identifiés. Le nombre de messages collectés pour chaque auteur est de 30. Chaque message comprend le style linguistique des auteurs tels que l'utilisation de chiffres, la ponctuation, les mots et les phrases fréquentes.

Sun et al. (2012) utilisent ce corpus pour mettre au point une méthode stylométrique visant à lutter contre la cybercriminalité. Ils utilisent des n-grammes de différentes longueurs pour représenter les textes, puis un premier filtrage sur la fréquence est effectué. Les n-grammes restants sont pondérés par le gain d'information puis une sélection de variable est opérée avec la méthode IGAE qui utilise un modèle « wrapper » à base d'algorithme génétique incorporant une heuristique également basée sur le gain d'information. Ils obtiennent une précision de 80,74% sur les 50 auteurs, et, 94,32% sur un sous-ensemble de 20 auteurs.

5. Les résultats

5.1. Deft'05

Les meilleurs résultats obtenus par la combinaison de notre méthode de maximisation des traits (FMC) avec un classifieur de type réseaux bayésien (BN) donnent une valeur de pertinence de 99.999% pour la tâche 1. Comme le montre la matrice de confusion présentée au tableau 6, nous n'avons que 12 erreurs, contre 16850 env. pour la meilleure approche antécédente (El-Bèze et al. 2005). De plus, contrairement aux approches antécédentes, les erreurs ne sont pas bilatérales. Mitterrand est confondu 12 fois avec Chirac, mais Chirac n'est jamais confondu avec Mitterrand. Pour obtenir ces résultats nous n'avons appliqué aucun traitement linguistique : il n'y a pas eu d'opération de lemmatisation, à part bien sûr une tokenisation des textes ; les « mots vides » ont été conservés et se montrent utiles pour l'analyse.

a	b	
73255	0	a = Chirac
12	11308	b = Mitterand

Tab 6. : Matrice de confusion pour le corpus Chirac-Mitterand (classification BN).

Le tableau 7 fait apparaître les expressions de plus fort contraste dans les discours de chaque protagoniste. Dans le cas de Mitterand, on retrouve bien les types de traits linguistiques mis en évidence par Habert (Habert et al. 2000), mais le panel de traits isolés est nettement plus exhaustif. Le discours de Mitterrand semble également marqué par des connotations humanistes comme l'illustre la suite de fort contraste « gens, assez, capables ». Le contraste des traits dominants est nettement plus prononcé pour le cas de Chirac et ces traits représentent en grande majorité des substantifs, matérialisant un discours plus clairement établi et basé sur des valeurs stables. Selon cette analyse, certain substantifs, comme le substantif « Asie », bien classés, peuvent sembler hors contrôle. La présence de ce dernier n'est pas aberrante, au contraire, puisqu'il correspond à une appétence personnelle bien connue de Chirac.

5.2. Amazon

La tâche consiste ici à reconnaitre l'auteur d'un message. Chaque message comprend le style linguistique des auteurs tels que l'utilisation de chiffres, la ponctuation, les mots et les phrases fréquentes. Pour cette raison, tous les mots, y compris les signes mentionnés ci-dessus, sont conservés dans cette base de données et l'espace de description résultant comprend 10000 mots.

Mitterand			Chirac	
Contraste	Terme		Contraste	Terme
1.88	douze		1.93	partenariat
1.85	est-ce		1.86	dynamisme
1.80	eh		1.81	exigence
1.79	quoi		1.78	compatriotes
1.78	-		1.77	vision
1.76	gens		1.77	honneur
1.75	assez		1.76	asie
1.74	capables		1.76	efficacité
1.72	penser		1.75	saluer
1.70	bref		1.74	soutien
1.69	puisque		1.74	renforcer
1.67	on		1.72	concitoyens
1.66	étais		1.71	réforme
1.62	parle		1.70	devons
1.62	fallait		1.70	engagement
1.60	simplement		1.69	estime
1.59	entendu		1.67	titre
1.58	suite		1.67	pleinement
1.57	peut-être		1.66	cœur
1.57	espère		1.66	ambition
1.56	parlé		1.65	santé
1.55	dis		1.64	stabilité
1.55	cela		1.63	amitié
1.54	existe		1.63	accueil
1.54	façon		1.62	publics

Tab 7. : Expressions les plus contrastées dans les discours de Chirac et Mitterrand.

Les résultats du tableau 8 montrent que dans le cas d'une tache de classification supervisée, en utilisant ici un algorithme bayésien multinomial (MNB), notre sélection de variables permet à nouveau d'aboutir à un résultat d'une précision très élevée. Nous avons seulement 3 messages mal classés alors que les résultats dans la littérature en donnent une moyenne de 378.

La méthode présente des résultats très supérieurs à l'état de l'art. En effet, en comparaison avec l'approche de référence de Sun et al. (2012), les résultats obtenus sur l'ensemble des classes couvertes par les 50 auteurs du corpus s'avèrent très supérieurs à ceux obtenus sur

un sous-ensemble réduit à 20 classes par lesdits auteurs (voir section 4.2). L'extraction des n-grammes de plus fort contrastes, présentés au tableau 9 sur une sélection de plusieurs classes, met en évidence que le fait que méthode permet, de manière additionnelle à la classification proprement dite, de caractériser les contributeurs dont le style est le plus spécifique. La méthode met clairement en évidence l'importance de la combinaison de marqueurs syntaxiques propres aux contributeurs et de contenu informatif spécifique dans la caractérisation des contributeurs. Ces deux types d'éléments sont parallèlement extraits par la méthode proposée.

		TP (R)	FP	P	F	ROC	TP Incr
Amazon	-	0.748	0.05	0.782	0.748	0.981	
	FMC	0.998	0.001	0.998	0.998	1	+33%

Tab 8. : Résultats de classification après sélection de variable FMC (classification MNB).

Ashbacher	Chacra	Chell	Cutey	Engineer
4.24 children	12.53 ..	5.38 hon	7.19 and_you	15.62 arge
3.82 _bas	12.43 ...	4.08 Amazon	6.33 highly	15.11 cel
3.45 bas	6.40 thus	3.68 az	5.87 really	14.58 arg
3.38 re_is	5.02 We	3.64 pho	5.87 reall	12.94 _ce
3.31 ines	5.02 /	3.29 nes_	5.73 eall	11.49 lls
3.26 _he_	4.87 and_thus	3.23 -s	5.73 eally	11.36 rge
3.17 _chil	4.66 used	3.10)._	5.21 songs	10.91 char
3.15 _chi	4.48 5	3.07).	5.05 fun	10.11 pac
3.06 chil	4.32 our	3.02 f_a_	4.81 love	9.59 lls_
3.05 dre	4.04 bel	2.93 of_a	4.70 it._	9.20 00
Janson	**Mark**	**McKee**	**Power**	**Sherwin**
6.97 art	7.33 track	12.28 oto	7.66 helpful	7.45 ;
5.87 omi	5.02 price	8.24 _pho	4.22 tern	6.85 ;_
4.98 horror	4.69 clea	7.13 _ph	3.93 e,_an	5.36 album
4.70 orr	4.44 _clea	6.83 photos	3.88 orie	4.80 tm
4.19 stories	4.16 _cle	6.32 pho	3.86 ,_y	4.59 Christmas
4.14 orie	4.10 music	5.25 ogr	3.80 may_	4.34 !_
3.96 ries_	4.03 So	4.87 photo	3.79 ories	3.94 rist
3.91 ories	3.94 nj	4.69 a_good	3.73 own	3.92 CD
3.91 cs	3.72 musi	4.19 ph	3.72 stories	3.77 !
3.78 stori	3.63 nice	3.77 op_	3.72 _may_	3.72 _very

Tab 9. : Description des classes à partir des n-grammes les plus contrastés (extrait).

6. Conclusion

L'objectif principal de cette communication était d'illustrer à la fois l'intérêt et l'avantage de la métrique de maximisation des traits que nous avons développée récemment pour l'analyse des données textuelles. Par l'intermédiaire d'une méthode de sélection de variables s'appuyant sur cette métrique, nous avons montré qu'il était possible de confirmer son efficacité sur des corpus textuels de référence de grande taille, hébergeant des données multi-classes, bruitées et/ou mal balancées, et connus pour poser des problèmes aux méthodes usuelles de classification et de sélection de variables, comme celles basées sur la métrique du Chi2.

Diverses expériences menées sur ces derniers corpus ont montré qu'elle permettait d'améliorer les performances des classifieurs dans un tel contexte, tout en mettant l'accent sur les classifieurs les plus flexibles et les moins gourmands en temps de calcul, comme les classifieurs bayésiens. Nous avons également montré que plus le nombre de classes examinées était important, plus les résultats obtenus s'avéraient supérieurs aux approches de référence.

De manière additionnelle, nous avons montré comment cette méthode permettait d'identifier et de visualiser de manière pertinente la pluridisciplinarité et les champs pluridisciplinaires dans les corpus, à travers l'extraction des traits caractéristiques des champs concernés.

Un autre avantage de la méthode présentée dans cette communication est qu'il s'agit d'une approche sans paramètres qui s'appuie sur un schéma simple d'extraction de variables ; elle peut donc être utilisée dans de nombreux contextes, comme dans ceux de l'apprentissage incrémental ou semi-supervisé, ou encore, dans celui de l'apprentissage numérique en général. Une autre perspective intéressante serait d'adapter cette technique au domaine de l'exploration de textes afin d'enrichir des ontologies et des lexiques grâce à l'exploitation à grande échelle des corpus existants.

Références

Alphonse É., Amrani A., Azé J., Heitz T., Mezaour A.-D. et Roche M. (2005). Préparation des données et analyse des résultats de DEFT'05, in *actes de la conférence « Traitement Automatique des Langues Naturelles » (TALN 2005) - Atelier DEFT'05*, 2005, vol. 2, pp. 99-111.

Apte C., Damerau F. et Weiss S.M. (1998). Text mining with decision rules and decision trees, in *Proceedings of the Conference on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web*.

Bache, K. et Lichman M. (2013). *UCI machine learning repository* (http://archive.ics.uci.edu/ml). University of California, School of Information and Computer Science, Irvine, CA, USA.

Bolón-Canedo, V., N. Sánchez-Maroño & A. Alonso-Betanzos (2012). A Review of Feature Selection Methods on Synthetic Data, in *Knowledge and Information Systems* (March 1, 2012): 1-37.

Choi F.Y.Y. (2000). Advances in domain independent linear text segmentation, in *Proceeding of NAACL-00*, pp. 26–33.

Cohen, A.M. et Hersh W.R. (2005). A survey of current work in biomedical text mining, in *Briefings in Bioinformatics* 6, pp. 57-71, 2005.

Cormack G.V. et Lynam T.R. (2007). Online supervised spam filter evaluation, in *ACM Transactions on Information Systems*, 25(3):11.

Daviet, H. (2009). Class-Add, une procédure de sélection de variables basée sur une troncature k-additive de l'information mutuelle et sur une classification ascendante hiérarchique en prétraitement. PhD, Université de Nantes, France.

Durant K. et Smith M. (2007). Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection, in *Advances in Web Mining and Web Usage Analysis: 8th International Workshop on Knowledge Discovery on the Web, Webkdd 2006*, pp. 187–206.

Falk, I., Lamirel J.-C., Gardent C., (2012). Classifying French Verbs Using French and English Lexical Resources, in *International Conference on Computational Linguistic (ACL 2012)*.

Guyon, I., Elisse A. (2003). An introduction to variable and feature selection, in *The Journal of Machine Learning Research*, 3:1157-1182.

Habert B., Illouz G., Lafon P., Fleury S., Folch H., Heiden S. et Prévost S. (2000). Profilage de textes : cadre de travail et expérience, in *Proc. of JADT'2000 (5ièmes journées internationales d'Analyse Statistique des Données Textuelles)*.

Hillard D., Purpura S. et Wilkerson J. (2007). An active learning framework for classifying political text, in *Annual Meeting of the Midwest Political Science Association*.

El-Bèze M., Torres-Moreno J.-M. et Béchet F. (2005). Peut-on rendre automatiquement à César ce qui lui appartient ? Application au jeu du Chirand-Mitterrac, in *Actes de la conférence « Traitement Automatique des Langues Naturelles » (TALN 2005) - Atelier DEFT'05*, 2005, vol. 2, pp. 125-134.

Hajlaoui K., Cuxac P., Lamirel J.C., François C. (2012). Enhancing patent expertise through automatic matching with scientific papers, in *Discovery Science*, LNCS 7569, pp. 299-312.

Ladha L., Deepa T. (2011). Feature selection methods and algorithms, in *International Journal on Computer Science and Engineering*, 3(5): 1787-1797.

Iyer R., Lewis D., Schapire R., Singer Y. et Singhal A. (2000). Boosting for document routing, in *Proceedings of the Ninth International Conference on Information and Knowledge Management*.

Jardino M. (2006). Identification des auteurs de textes court avec des n-grammes de caractères, in *Proc. of JADT'2006 (8èmes journées internationales d'Analyse Statistique des Données Textuelles)*, vol. 2, pp. 543-549.

Joachims, T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, in *Proceedings of ICML-97, 14th International Conference on Machine Learning*,

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features, in *Proceedings of the European conference on Machine learning*, pp.137-142.

Labadié A., Romero Y. et Sitbon L. (2005). Segmentation et classification : deux politiques complémentaires, in *Actes de la conférence « Traitement Automatique des Langues Naturelles » (TALN 2005) - Atelier DEFT'05*, 2005, vol. 2, pp. 183-192.

Lamirel, J.-C., Al Shehabi S., François C. et Hoffmann M. (2004). New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping, in *Scientometrics*, 60(3).

Lamirel J.-C., Ta A.P., Attik M. (2008). Novel labeling strategies for hierarchical representation of multidimensional data analysis results, in *Proceedings of IASTED International Conference on Artificial Intelligence and Applications (AIA)*.

Lamirel, J.-C., Mall R., Mall R., Cuxac P., Safi G., (2011). Variations to incremental growing neural gas algorithm based on label maximization, in *Proceedings of IJCNN 2011*.

Lamirel J.-C., Cuxac P., Chivukula A.S., Hajlaoui K. (2013). A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data, in *QIMIE 2013: 3nd International PAKDD Workshop on Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models*.

Lewis D.D. (1992). An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task, in *ACM 15th Ann Int'l SIGIR'92*, pp. 37-50.

Pang B. et Lee L. (2008). Opinion mining and sentiment analysis, in *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

Pierron L., Durkal C. et Chevalier J.-B. (2005). Classification, combinaison et regroupements pour séparer les discours de Mitterrand de ceux de Chirac, in *Actes de la conférence « Traitement Automatique des Langues Naturelles » (TALN 2005) - Atelier DEFT'05*, 2005, vol. 2, pp. 165-173.

Purpura S. et Hillard D. (2006) Automated classification of congressional legislation, in *Proceedings of the International Conference on Digital Government Research*, pp. 219–225.

Rigouste L., Cappé O. et Yvon F. (2005). Modèle de mélange multi-thématique pour la Fouille de Textes, in *Actes de la conférence « Traitement Automatique des Langues Naturelles » (TALN 2005) - Atelier DEFT'05*, 2005, vol. 2, pp. 193-202.

Schapire R., Singer Y. et Singhal A. (1998). Boosting and Rocchio applied to text filtering, in in *ACM 21st Ann Int'l SIGIR'98*.

Schütze, H., Hull D.A. et Pedersen J.O. (1995). A Comparison of Classifiers and Document Representations for the Routing Problem, in *ACM 18th Ann Int'l SIGIR'95*, pp. 229-337.

Sun J., Yang Z., Liu S., et Wang P. (2012). Applying Stylometric Analysis Techniques to Counter Anonymity in Cyberspace, in *Journal of Networks*, vol. 7, n° 2, févr. 2012.

Wiener E., Pedersen J.O et Weigend A. S. (1995). A Neural Network Approach to Topic Spotting, in *Symposium on Document Analysis and Information Retrieval*, pp. 317-332.

Yang Y. et Chute C.G. (1994). An example based mapping method for text categorization and retrieval, in *ACM Trans. Inform. Syst.*, 12: 252-277, 1994.