Etude comparative de stratégies de sélection de prédicteurs pour l'attribution d'auteur

Jacques Savoy

Institut d'informatique, Université de Neuchâtel
rue Emile Argand 11, 2000 Neuchâtel (Suisse)

Jacques.Savoy@unine.ch

RÉSUMÉ. *L'attribution d'auteur peut être vue comme une tâche en catégorisation de textes qui se subdivise en deux étapes. D'abord nous devons sélectionner les mots les plus discriminants puis appliquer un modèle de classification. Afin de bien choisir les meilleurs termes, nous avons évalué sept fonctions de sélection dont l'information mutuelle ponctuelle, le gain d'information, le rapport de cotes, le χ^2 ou le coefficient de corrélation. Nous avons également retenu deux stratégies de sélection proposées dans le cadre d'attribution d'auteur. Afin de comparer ces méthodes, nous avons repris un corpus de 5 408 articles de presse (Glasgow Herald) écrits par vingt journalistes différents. Basé sur la performance obtenue par la méthode de divergence KLD (Zhao & Zobel, 2007) et Delta (Burrows, 2002), nous remarquons que des stratégies simples proposent des résultats aussi performants que des approches plus complexes.*

ABSTRACT. *The authorship attribution problem can be viewed as a categorization problem. To determine the most effective features to discriminate between different writers (or categories), we have evaluated seven feature selection functions (e.g., pointwise mutual information, information gain, odds ratio, χ^2, or correlation coefficient). We have also considered two selection functions proposed in the context of authorship attribution. To compare these approaches, we have selected a newspaper corpus (Glasgow Herald) composed of 5,408 articles written by twenty columnists. Using the KLD (Zhao & Zobel, 2007) and the Delta (Burrows, 2002) attribution scheme, we found that some simple selection functions tend to produce results comparable to more complex ones.*

MOTS-CLÉS : *Sélection de prédicteurs, attribution d'auteur, catégorisation de textes.*

KEYWORDS: *Feature selection, authorship attribution, text categorization.*

1. Introduction

L'attribution d'auteur cherche à déterminer l'auteur d'un écrit anonyme ou dont l'attribution reste incertaine (Love, 2002). Comme objet d'étude, on rencontre des lettres, des œuvres littéraires (voir le débat Molière-Corneille (Labbé, 2009)), ou des fragments de celles-ci (pour déterminer les passages vraiment écrits par Shakespeare (Craig & Kinney, 2009)) voire des discours politiques (T. Sorensen dans l'ombre du

Président Kennedy (Carpenter & Seltzer, 1970) (Monière & Labbé, 2006)) ou des courriels.

Afin de résoudre cette question, une première famille d'approches désire recourir à un nombre limité de mots fonctionnels fréquents afin de cerner le style de l'auteur de manière indépendante des thèmes abordés. Dans un second paradigme, l'attribution d'auteur peut être analysée sous l'angle de la catégorisation de textes (Sebastiani, 2002), (Manning *et al*., 2008) dans laquelle chaque auteur potentiel correspond à une catégorie. Dans cette optique, les textes doivent être représentés par des caractéristiques (mots, *n*-grammes de caractères, lemmes, parties du discours, brèves séquences de ces dernières, etc.) ayant la capacité de discriminer entre les diverses catégories. Sur ces représentations, on entraine un classifieur afin qu'il puisse détecter les particularités propres à chaque auteur.

Proposer de résoudre automatiquement l'attribution d'auteur en recourant à des techniques de catégorisation automatique implique l'idée que les deux domaines partagent des caractéristiques communes. En effet, dans les deux cas les textes doivent être représentés en s'appuyant sur les mots présents, leurs fréquences, voire leurs positions. De même, la taille très importante du vocabulaire nécessite un élagage et une sélection des termes les plus adéquats pour distinguer les diverses catégories sous-jacentes. Toutefois, l'attribution d'auteur possède ses traits propres. Ainsi, la distinction entre auteurs devrait s'appuyer sur les différences de style et, dans cette perspective, la prise en compte de la ponctuation ou des mots outils s'avère pertinente. Enfin, le recours à un séparateur général s'avère, pour certains auteurs, peu efficace comparé à une règle de décision plus simple fondée uniquement sur un nombre restreint de formes très fréquentes.

L'objectif de cet article est de comparer les diverses stratégies de sélection des prédicteurs en attribution d'auteur afin de déterminer si la spécificité de cette tâche permet de baser une décision uniquement sur un nombre restreint de mots très fréquents. De plus, nous souhaitons connaître la variation de l'efficience par la prise en compte d'un nombre plus important de termes. Dans la suite de cet article, nous présenterons les principales stratégies suggérées dans la sélection des vocables pour l'attribution d'auteur (section 2). La troisième section expose les grandes lignes du corpus utilisé dans nos expériences. La quatrième section décrit quelques méthodes utilisées pour la sélection de prédicteurs. La cinquième section présente deux modèles de classification performants en attribution d'auteur et la sixième résume l'évaluation des fonctions de sélection avec nos deux séparateurs. Finalement, une conclusion dresse les principales contributions de cette étude.

2. État des connaissances

Afin de proposer une solution automatique en attribution d'auteur (Juola, 2006), les premières études ont cherché à définir une mesure stylométrique devant être constante pour un auteur et différente d'un écrivain à l'autre (Holmes, 1998). Ainsi,

on a proposé de tenir compte de la longueur moyenne des mots ou des phrases, du nombre moyen de syllabes par mots, voire de la taille du vocabulaire V (notée |V|) par rapport à la longueur du document. Comme alternative, on a proposé la valeur R = |V| / sqrt(*n*)) de Guiraud avec |V| indiquant la taille du vocabulaire, le rapport entre le nombre de *hapax legomena* (notée V_1) et la taille du vocabulaire (soit $|V_1| / |V|$), ou le rapport entre le nombre de mots apparaissant deux fois (noté $|V_2|$) et la taille du vocabulaire (Sichel, 1975). Toutefois, ces mesures ont l'inconvénient d'être assez instables (Baayen, 2008), en particulier face à des documents relativement courts (de taille inférieure à mille mots). De plus, le genre (poésie, pièce de théâtre, roman, texte en vers ou en prose) influence de telles mesures.

Afin de fonder les décisions d'attribution sur le vocabulaire, Mosteller & Wallace (1964) proposent de sélectionner de manière semi-automatique les vocables les plus pertinents. Cette étude met en lumière l'importance des mots fréquents et, en particulier, des mots fonctionnels (déterminants, prépositions, conjonctions, pronoms et quelques auxiliaires). Par exemple, les auteurs remarquent que le terme *language* est utilisé deux fois par Hamilton mais dix fois par Madison. Dans ce raisonnement, on admet que la fréquence d'apparition de certains mots ne sont pas sous le contrôle conscient de l'auteur et qu'ils varient d'une personne à l'autre.

En poursuivant cette voie, Burrows (2002) propose de sélectionner les mots pouvant refléter le style d'un auteur et qui soient indépendants du thème traité. Dans cette perspective, le critère de sélection retenu se limite à la fréquence d'occurrence. Ainsi le vocabulaire à retenir comprendra les 50 à 150 vocables les plus fréquents, ensemble comprenant une forte proportion de mots fonctionnels. Ce seuil sera repoussé à 800 (Hoover, 2004) puis à 4 000 (Hoover, 2007) avec l'inclusion de mots lexicaux fréquents (noms, adjectifs, adverbes et verbes).

Les études menées par Zhao & Zobel (2005, 2007) proposent de définir *a priori* les vocables à retenir. Dans ce cas, on retient essentiellement les mots fonctionnels en ignorant les mots lexicaux liés aux thématiques. Pour la langue anglaise, ces auteurs suggèrent une liste de 363 formes, un ensemble correspondant au contenu d'une liste de mots outils d'un moteur de recherche.

Finalement, d'autres auteurs proposent de s'appuyer sur des techniques développées dans le cadre de la catégorisation thématique (Stamatatos, 2009). Dans cette perspective, nous devons d'abord sélectionner les termes possédant le meilleur pouvoir discriminant puis entraîner un séparateur. Dans cette étude, nous nous intéressons à la première phase. Dans ce cadre, l'étude comparative de Yang & Pedersen (1999) évalue six mesures de sélection, sur deux corpus et à l'aide de deux classifieurs (*k-Nearest Neighbors* et *Linear Least Squares Fit*). Leurs résultats indiquent qu'un élagage basé sur la fréquence documentaire (*df*) apporte des résultats similaires à des méthodes plus complexes basées sur le gain d'information (nommé aussi *expected mutual information*) ou du χ^2. Pour Sebastiani (2002), le rapport de cotes (*odds ratio*) et la métrique du χ^2 permettent d'obtenir généralement les meilleures performances.

Toutefois, une différence importante persiste entre l'attribution d'auteur et la catégorisation thématique. En effet, dans cette dernière, on propose d'éliminer les mots très fréquents et peu ou pas porteurs de sens (Yang, 1999) (Sebastiani, 2002), tandis que ces derniers sont valorisés comme marqueurs de style. Enfin, des études plus récentes en attribution d'auteur tendent à se fonder sur d'autres éléments que le lexique comme la présence d'une signature, la mise en pages, le type et la fréquence des césures ou l'usage d'étiquettes HTML (Zheng *et al*., 2006). Avec l'adjonction de ces caractéristiques augmentant l'espace de représentation, la nécessité d'une bonne stratégie de sélection se trouve renforcée.

3. Corpus d'évaluation

Grâce à des collections tests, nous pouvons évaluer et comparer divers représentations et classifieurs. Contrairement à la catégorisation automatique, les études en attribution d'auteur disposent d'un nombre restreint de corpus. De plus, les corpus disponibles comprennent un nombre limité de documents et seulement quelques auteurs potentiels (par exemple, les *Federalist Papers* (Mosteller & Wallace, 1964) comprennent 85 articles et la paternité de 12 d'entre eux demeure incertaine (on hésite essentiellement entre deux auteurs possibles)).

Désirant fonder nos conclusions sur une base plus large et au moyen d'une collection stable et facilement accessible, nous avons sélectionné un sous-ensemble de la collection CLEF- 2003 (Peters *et al*., 2004). Cette partie comprend les articles publiés durant l'année 1995 dans le journal *Glasgow Herald*. Si le corpus complet compte 56 472 documents, nous ne connaissons le ou les auteur(s) que pour 28 687 d'entre eux. De ce dernier sous-ensemble, nous avons sélectionné les articles rédigés par un seul auteur et écarté les journalistes ayant écrit peu d'articles durant l'année 1995. Finalement, nous avons obtenu un corpus de 5 408 articles écrits par vingt auteurs différents.

Dans le tableau 1 nous avons indiqué le nom des journalistes, le thème principal correspond à chaque auteur, puis le nombre d'articles rédigés. On constate que le nombre d'articles par journaliste varie fortement entre le minimum de 30 (J. Fowler) et le maximum de 433 (A. Wilson). En dernière colonne, nous avons indiqué la longueur moyenne (en nombre de mots) des articles rédigés, subdivisés par auteur. Sur cette base, on constate que cette moyenne varie fortement entre auteurs, avec une valeur minimale de 452 (A. Wilson) jusqu'à un maximum de 1 301 (J. Davidson).

Si nous attribuons de manière aléatoire entre les vingt auteurs chaque document, nous obtiendrons un taux de réussite proche des 5 %. Si nous tenons compte du fait que les vingt journalistes n'ont pas été le même nombre de documents, nous pouvons choisir systématiquement l'auteur du plus grand nombre d'articles (A. Wilson). Dans ce cas de figure, la taux de réussite s'élèverait à 8 % (433 / 5408). Cette valeur limite représente la performance minimale d'un système d'attribution.

Les séparateurs étudiés vont nous permettre d'obtenir des performances supérieures en s'appuyant sur une représentation adéquate des divers textes et profil d'auteur.

Afin de représenter un article, nous devons nous fonder sur des termes relativement fréquents. Ainsi, l'apparition d'un mot usité une seule fois dans un corpus (*hapax legomena*) doit être ignorée. Cette technique d'élagage permet de réduire le vocabulaire des articles du *Glasgow Herald* de 79 220 vocables à 45 402 (diminution relative de 42,7 %). Ensuite, nous avons éliminé les termes présents uniquement chez l'un des journalistes considérés. Certes la « signature » d'une personne peut se relever par un usage exclusif de certaines formes (par exemple, la *chienlit* de C. de Gaulle ou le *abracadabrantesque* de J. Chirac). Par contre, le système peut également être plus facilement trompé par l'emploi d'une telle forme. L'application de cette règle réduit encore notre vocabulaire dont la taille s'élève à 36 773 vocables (soit 46,4 % du volume initial). Enfin, afin de garantir une représentation des écrits se basant plus sur des éléments de style, nous avons décidé de ne retenir que les vocables apparaissant deux fois au moins dans un article. Après ce dernier élagage, la taille du vocabulaire possible comprendra 10 994 entrées, soit 13,9 % de la taille initiale.

La question que l'on désire résoudre est de savoir quelle stratégie de sélection permettra d'extraire de cet ensemble relativement important de 10 994 termes, un nombre plus restreint de prédicteurs efficients.

	Nom	Thème	Nombre d'articles	Longueur moyenne
1	Davidson Julie	arts & cinéma	57	1 310
2	Douglas Derek	sports	410	808
3	Fowler John	arts & cinéma	30	890
4	Gallacher Ken	sports	408	727
5	Gillon Doug	sports	368	713
6	Johnstone Anne	politique	72	1 258
7	McConnell Ian	business	374	455
9	McLean Jack	social	118	1 008
9	Paul Ian	sports	418	842
11	Reeves Nicola	business	370	531
11	Russell William	arts & cinéma	291	1 019
12	Shields Tom	politique	173	1 001
13	Sims Christopher	business	390	471
14	Smith Ken	social	212	616
15	Smith Graeme	social	329	520
16	Traynor James	sports	339	983
17	Trotter Stuart	politique	336	666
18	Wilson Andrew	business	433	452
19	Wishart Ruth	politique	72	1 137
20	Young Alf	business	208	1 013

Tableau 1. Répartition des articles sélectionnés par journaliste (*Glasgow Herald*, 5 408 articles)

4. Méthodes de sélection

Afin de déterminer les termes les plus pertinents, nous disposons de plusieurs fonctions de sélection évaluant le pouvoir discriminant du terme t_k pour la catégorie c_i, avec $i = 1, 2, \ldots, |C|$. Afin de définir de telles fonctions, nous procédons sur la base d'une table de contingence pour chaque couple (t_k, c_i) dont un exemple est repris dans le tableau 2. Dans ce dernier, la valeur a indique le nombre d'assignation du terme t_k à la catégorie c_i. Si l'on considère toutes les autres catégories (ensemble désigné par $-c_i$), le terme t_k apparaît b fois. Sur l'ensemble du corpus, le terme t_k apparaît donc $a+b$ fois. Le nombre c indique le nombre d'instances de la catégorie c_i qui ne possède pas le terme t_k. La valeur $a+c$ indique le nombre de documents appartement à la catégorie c_i.

	Catégorie c_i	Catégorie $-c_i$	
Terme t_k	a	b	$a + b$
Autres termes $-t_k$	c	d	$c + d$
	$a + c$	$b + d$	$n = a+b+c+d$

Tableau 2. *Exemple d'une table de contingence pour le terme t_k et la catégorie c_i*

Afin de mesurer le degré d'association entre un terme donné t_k et une catégorie c_i, on peut se baser uniquement sur la probabilité Prob[c_i | t_k], une approche nommée aussi facteur d'association (ou DIA, *Darmstadt Indexing Approach* (Fuhr *et al.*, 1991)). Basé sur le tableau 2, nous pouvons estimer Prob[c_i | t_k] = a / (a+b). Afin d'alléger la présentation, nous avons regroupé les formules et estimations dans l'annexe.

Comme seconde possibilité, nous pouvons recourir à la mesure de l'information mutuelle ponctuelle (IMP, *pointwise mutual information*) (Church & Hanks, 1987). En présence d'indépendance entre la catégorie c_i et le terme t_k, la valeur IMP tend vers 0. Une association positive entre la catégorie et le terme se signale par une valeur positive, et inversement par une valeur négative. Cette mesure tend toutefois à favoriser les termes plutôt rares (Dunnig, 1993).

Comme troisième approche, nous avons retenu le rapport de cotes OR (*odds ratio*) (Caropreso *et al.*, 2001) donnant toujours une valeur positive. Une association entre le terme t_k et la catégorie c_i s'indique par une valeur plus grande que l'unité, tandis qu'une opposition se signale par une valeur proche de zéro.

Comme quatrième mesure, le gain d'information (GI, nommé également *expected mutual information*) a été retenu. Cette mesure sera élevée s'il existe une association entre le terme sous-jacent et la catégorie concernée. L'absence d'un pouvoir discriminant pour ce terme et cette catégorie s'indiquera par une faible valeur positive. Suivant une règle d'interprétation similaire, nous avons également considéré la mesure de chi-carrée $\chi^2(t_k, c_i)$ (Gavalotti *et al.*, 2000). Dans ce cas, une valeur proche de zéro indique une indépendance entre le terme et la catégorie concernée. Une association ou opposition forte sera indiquée par une valeur élevée.

Dérivé de la mesure χ^2, le coefficient de corrélation $CC(t_k, c_i)$ (Ng *et al*. 1997) correspond à la sixième fonction retenue. Dans ce cas, une association positive se signale par une valeur positive, tandis qu'une opposition sera signalée par une valeur négative. Une valeur proche de zéro symbolise l'absence de lien entre le terme et la catégorie. Finalement et suivant la même interprétation, le coefficient GSS peut également servir à sélectionner les meilleurs termes (Gavalotti *et al*., 2000).

En plus de ces sept fonctions de sélection, nous pouvons également retenir la fréquence documentaire (*df*) indiquant le nombre de documents indexés par le terme t_k. Cette stratégie apporte de bons résultats (Yang & Pedersen, 1997) et a déjà été proposée en attribution d'auteur (Grieve, 2007). De plus, le style d'un écrivain peut se signaler par l'emploi de mots fonctionnels ou par l'usage fréquent de certaines formes. Dans cette perspective, nous pourrions ainsi suivre Burrows (2002) et recourir à la fréquence d'occurrence absolue (*tfa*) pour sélectionner les termes les plus utiles et pour distinguer les divers styles.

En appliquant l'une des fonctions décrites ci-dessus, nous obtenons une valeur d'utilité locale, notée $f(t_k, c_i)$, pour chaque terme t_k et catégorie c_i. En présence d'une catégorisation binaire, cette fonction suffit pour définir une valeur sélective à chaque terme. En règle générale, nous devons faire face à un nombre plus élevé de catégories (ou auteurs dans notre cas). Afin de comparer de manière globale les termes entre eux, nous devons agréger les valeurs locales sur l'ensemble des $|C|$ catégories. Pour définir une telle valeur d'utilité globale d'un terme t_k (notée $U(t_k)$), on peut calculer le maximum sur toutes les catégories ou la somme pondérée (en fonction de l'importance de chaque catégorie) comme l'indique l'équation 1.

$$U(t_k) = Max_i f(t_k, c_i), \qquad U(t_k) = \sum_{i=1}^{|C|} \text{Prob}[c_i] \cdot f(t_k, c_i) \tag{1}$$

Afin de sectionner les *m* termes les plus adaptés à discriminer entre les catégories, nous prendrons les *m* termes ayant les valeurs d'utilité $U(t_k)$ les plus élevées selon la formule d'agrégation (maximum ou somme pondérée).

5. Méthodes d'attribution

Comme méthode d'attribution d'un texte à un auteur, nous avons retenu l'approche proposée par Zhao & Zobel (2005, 2007). Ces derniers suggèrent de mesurer la distance entre le profil d'un auteur A_j (concaténation de tous ses écrits) et un texte requête (noté Q) en utilisant la divergence Kullback-Leibler (KLD) (nommée aussi entropie relative (Maning & Schütze, 1999)). Cette mesure est exprimée dans l'équation 2 dans laquelle $Prob_q[t_k]$ et $Prob_{aj}(t_k)$ indiquent la probabilité d'occurrence du terme t_k dans la requête ou le j^e profil d'auteur A_j. Lors du calcul, nous imposons que $0 \cdot \log_2[0/p] = 0$, et $p \cdot \log_2[p/0] = \infty$.

$$KLD(Q \,||\, A_j) = \sum_{k=1}^{m} \text{Prob}_q[t_k] \cdot \log_2 \left[\frac{\text{Prob}_q[t_k]}{\text{Prob}_{aj}[t_k]} \right] \quad (2)$$

Lorsque deux distributions sont identiques, la valeur KLD sera nulle. Dans tous les autres cas, la valeur retournée sera positive, et d'autant plus importante si la distance entre les distributions dérivées du document Q et du profil A_j est élevée.

Pour estimer les probabilités sous-jacentes, nous avons appliqué le principe du maximum de vraisemblance en estimant que $\text{Prob}[t_k] = tfa_k/n$, avec tfa_k indiquant la fréquence d'occurrence du terme et n la taille du document concerné. Cette estimation peut être lissée afin d'éliminer la présence de probabilités nulles (Manning & Schütze, 1999). Dans nos évaluations, nous avons adopté l'approche de Lidstone en estimant les probabilités par $(tfa_k+\lambda) / (n+\lambda \cdot |V|)$, avec $|V|$ indiquant la taille du vocabulaire retenue. Nous avons fixé la valeur du paramètre λ à 0,01 car cette dernière retourne la meilleure performance.

Comme seconde méthode d'attribution, nous avons retenu le modèle Delta (Burrows, 2002) mesurant la distance entre deux textes par des fréquences standardisées (score Z). Cette valeur est obtenue depuis la fréquence relative (notée tfr_{kj} pour le terme t_k dans le document D_j) par soustraction de la moyenne (notée $mean_k$) et division par l'écart-type (sd_k), moyenne et écart-type estimés en considérant le corpus sous-jacent (Hoover, 2004).

$$Z\ score(t_{kj}) = \frac{tfr_{kj} - mean_k}{sd_k} \quad (3)$$

Cette valeur est associée à chaque vocable retenu pour chaque document ou profil d'auteur. A l'aide de ces valeurs, on peut calculer la distance Delta Δ entre un document requête noté Q et un profil d'auteur noté A_j selon la formule 4.

$$\Delta(Q, A_j) = 1/m \cdot \sum_{k=1}^{m} \left| Z\ score(t_{kq}) - Z\ score(t_{kj}) \right| \quad (4)$$

Dans cette formulation, nous attachons la même importance à chaque terme t_k. Une différence importante entre Q et A_j apparaît lorsque, pour un vocable donné, les deux scores Z sont élevés et de signe opposé. A l'inverse, si le terme est usité avec la même fréquence relative dans les deux textes, la différence des scores Z sera faible, indiquant un rapprochement possible des deux textes. Finalement, si pour les m termes retenus les différences entre les scores Z demeurent faibles, la distance Δ résultante sera minime, indiquant que les deux textes sont probablement écrits par la même personne.

6. Evaluation

Avec notre corpus *Glasgow Herald* (5 408 articles, 20 auteurs), nous avons évalué l'approche KLD en utilisant les 363 mots définis a priori par Zhao & Zobel (2007). Cette liste contient essentiellement des mots fonctionnels (*the*, *in*, *but*, *not*, *am*, *of*, *can*, …), de même que des termes fréquents (*became*, *nothing*, …). Quelques entrées s'avèrent peu fréquentes (*howbeit*, *whereafter, whereupon*), indiquent le comportement attendu lors de la segmentation (*doesn, weren*) ou correspondent à un choix plus arbitraire (*indicate*, *missing*, *specifying*, *seemed*). Comme 19 mots n'apparaissent pas dans notre corpus, le nombre de mots réellement utilisés sera de 363 – 19 = 344.

Afin d'évaluer la performance de nos séparateurs, nous devons réserver des instances pour l'apprentissage et des exemples distincts pour le test. Pour respecter cette contrainte, nous pourrions adopter la validation croisée comme stratégie d'évaluation (Hastie *et al*. 2009). Dans le cas présent, nous avons choisi l'approche *leaving-one-out* attribuant toutes les instances, sauf une, pour l'entraînement et la dernière pour le test. Enfin, nous itérons cette démarche sur l'ensemble des 5 408 articles, chacun à tour de rôle est exclu de l'ensemble destiné à l'apprentissage.

En appliquant cette stratégie d'évaluation et sur la base des 344 termes définis a priori, le taux de réussite (*micro-average*) de l'approche KLD correspond à 70,8 %, valeur que nous avons indiquée en première ligne du tableau 3. Cette sélection faite manuellement et *a priori* peut être comparée aux neuf autres approches automatiques de sélection, avec la fonction agrégation maximum ou somme pondérée. Le tableau 3 redonne, pour chaque fonction de sélection, la meilleure combinaison du nombre de termes à sélectionner et la fonction d'agrégation.

	KLD		Delta	
	Paramètre	Perform.	Paramètre	Perform.
	344 mots	70,8 %	400	**63,7 %**
$df(t_k,c_i)$	1 500 / max	85,2 % †	300 / max	62,9 %
$tfa(t_k,c_i)$	2 000 / somme	**85,6 %** †	300 / somme	61,2 % †
$DIA(t_k,c_i)$	2 000 / somme	85,1 % †	150 / somme	58,3 % †
$\chi^2(t_k,c_i)$	5 000 / somme	84,4 % †	150 / max	38,7 % †
$GSS(t_k,c_i)$	2 000 / max	82,3 % †	150 / max	34,0 % †
$GI(t_k,c_i)$	3 000 / somme	84,6 % †	150 / max	35,4 % †
$CC(t_k,c_i)$	2 000 / somme	78,0 % †	3 000 / max	15,4 % †
$IMP(t_k,c_i)$	4 000 / somme	78,9 % †	2 000 / max	15,1 % †
$OR(t_k,c_i)$	4 000 / somme	64,7 % †	3 000 / max	12,6 % †

Tableau 3. *Évaluation des diverses stratégies de sélection avec les approches KLD (Zhao & Zobel, 2007) ou Delta (Burrows, 2002)*

Afin de comparaison, nous avons repris la meilleure performance de la méthode Delta (Burrows, 2002) qui s'obtient en considérant les 400 termes les plus fréquents dans le corpus. Le taux de réussite s'élève alors à 63,7 %. Dans la quatrième colonne du tableau 3, nous avons repris les neuf fonctions de sélection pour déterminer le nombre optimum de termes à retenir de même que la fonction d'agrégation avec la méthode Delta.

Les résultats obtenus dans ce tableau indiquent que les meilleures stratégies de sélection reposent sur des méthodes simples comme le DIA, la fréquence documentaire (*df*) ou d'occurrence (*tfa*). Ces deux dernières se rencontrent fréquemment dans les études empiriques en attribution d'auteur.

Comme deuxième choix, nous rencontrons la métrique du χ^2, la fonction GSS et le gain d'information (GI). On notera toutefois que dans le cadre du modèle KLD, la différence de performance avec les sélections simples (*df*, *tfa* ou DIA) ne s'avère pas très importante. Le coefficient de corrélation CC, le rapport de cotes (OR) ou l'information mutuelle ponctuelle (IMP) s'avèrent des choix peu intéressants, dans le cadre de l'attribution d'auteur pour le moins.

Afin de savoir si une différence de performance entre deux approches s'avère statistiquement significative, nous avons opté pour le test du signe (Conover, 1971), (Yang & Liu, 1999) (test bilatéral) avec un seuil de signification $\alpha = 1$ %. En appliquant ce test, l'hypothèse H_0 admet que les deux modèles possèdent des niveaux de performance similaire. Dans la table 3, nous avons retenu la première ligne comme modèle de référence et les différences de performance statistiquement significatives sont indiquées par une croix '†'. Comme on le constate, les performances obtenues après sélection des termes sont très souvent significativement différents du modèle de départ.

Au niveau du nombre de termes à retenir pour représenter les documents et le profil d'auteur, nous constatons que la méthode Delta nécessite un nombre restreint de mots (entre 150 et 400). Dans le cadre de ce modèle, la sélection des bons prédicteurs se limite à la fréquence documentaire (*df*) ou d'occurrence (*tfa*). Les autres méthodes de sélection tendent à pénaliser plus ou moins fortement la performance globale.

Pour l'approche KLD basée sur 344 mots, nous constatons que la prise en compte d'un nombre plus élevé (environ 1 500 à 3 000 termes) permet d'accroître de manière significative la performance (de 70,8 % à environ 85 %). De plus, diverses méthodes de sélection offrent des gains de performance assez similaire.

Afin de mieux comprendre les différences entre les méthodes de sélection, nous avons calculé le pourcentage de termes communs sélectionnés par deux fonctions de sélection. En nous limitant à la fonction d'agrégation somme et en faisant varier le nombre de termes entre 150 et 3 000, nous avons constaté que les fonctions DIA, *df* et *tfa* retournent, en moyenne, les ensembles de termes fortement similaires (entre 92 % à 100 % identique). De même, les ensembles de mots sélectionnés par les

fonctions CC et χ^2 sont très similaires, ce qui s'explique par le fait que la fonction CC est dérivée du calcul de la mesure χ^2. Il existe un rapprochement possible entre les fonctions GSS et GI dont les ensembles de termes sélectionnés disposent, en moyenne, d'un recouvrement de l'ordre de 77 %. Enfin, les fonctions OR et IMP ne se rapprochent clairement d'aucune autre, opérant des sélections fort distinctes.

7. Conclusion

Dans le cadre de cette communication, nous avons présenté l'attribution d'auteur comme une tâche particulière en catégorisation de textes. Dans ce cadre, la sélection des termes pouvant être discriminatoires entre les diverses catégories représente une composante centrale pour atteindre une bonne qualité de réponses.

Afin de pouvoir évaluer et comparer différentes fonctions de sélection, nous avons retenu sept fonctions ainsi que deux stratégies de sélection couramment usitées en attribution d'auteur. Comme méthode d'attribution, nous avons repris la divergence Kleiber-Leibner proposée par Zhao & Zobel (2005 ; 2007) ainsi que la règle Delta (Burrows, 2002), deux méthodes proposant de très bonnes performances.

Sur la base d'un corpus d'articles de presse (*Glasgow Herard*) comprenant 5 408 articles, écrits par vingt journalistes, nos évaluations indiquent que des stratégies de sélection basées sur la fréquence documentaire (*df*) ou d'occurrence (*tfa*) tendent à fournir de très bons résultats, comparables à la fonction DIA. Dans une deuxième classe de performance on retrouve la métrique du χ^2, la fonction GSS et celle du gain d'information (GI). L'emploi de l'information mutuelle ponctuelle (IMP), du coefficient de corrélation (CC) ou du rapport de cotes (OR) ne permettent pas d'apporter une sélection efficace des termes, dans le cadre de l'attribution d'auteur pour le moins.

Contrairement à l'étude de Yang & Pedersen (1997) conduite dans le cadre de la catégorisation thématique, les mesures de gain d'information (GI) ou χ^2 ne correspondent pas aux meilleures stratégies de sélection en attribution d'auteur. De même, Sebastiani (2002) indique que les meilleures fonctions de sélection sont le rapport des cotes avec l'opérateur d'agrégation somme (OR_{sum}) ou le GSS_{max}. Notre étude indique que dans le cadre de l'attribution d'auteur pour le moins, ces choix ne s'avèrent pas pertinents.

Remerciements

L'auteur tient à remercier les trois relecteurs anonymes pour leurs commentaires constructifs dans la rédaction de cette communication.

8. Bibliographie

Baayen R.H. *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge, Cambridge University Press, Cambridge, 2008.

Burrows J.F. « Delta: A measure of stylistic difference and a guide to likely authorship », *Literary and Linguistic Computing*, vol. 17, n° 3, 2002, p. 267-287.

Caropreso M.F., Matwin S. & Sebastiani F. « A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization », In A.G. Chin, *Text Databases and Document management: Theory and Practice*. Hershey, Idea 2001, p. 78-102.

Carpenter R.H. & Seltzer R.V. « On Nixon's Kennedy style », *Speaker and Gavel*, 7(41), 1970.

Church K.W. & Hanks P. « Word association norms, mutual information and lexicography », *Proceedings ACL*, 1989, p. 76-83.

Conover W.J. *Practical Nonparametric Statistics*, 2nd Ed., New York, John Wiley & Sons, 1971.

Craig H. & Kinney A.F. *Shakespeare, Computers, and the Mystery of Authorship*, Cambridge, Cambridge University Press, 2009.

Dunning T.E. « Accurate methods for the statistics of surprise and coincidence », *Computational Linguistics*, vol. 19, n° 1, 1993, p. 61-74.

Fuhr N., Hartmann S., Knorz G., Lustig G., Schwantner M. & Tzeras K. « AIR/X a rule-based multi-stage indexing system for large subject fields », *Proceedings RIAO*, 1991, p. 606-623.

Gavalotti L., Sebastiani F. & Simi M. « Experiments on the use of feature selection and negative evidence in automated text categorization », *Proceedings ECDL*, 2000, p. 59-68.

Grieve J. « Quantitative authorship attribution: An evaluation of techniques », *Literary and Linguistic Computing*, vol. 22, n° 3, 2007, p. 251-270.

Hastie T., Tibshirani R. & Friedman J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd Ed., New York, Springer, 2009.

Holmes D.I. « The evolution of stylometry in humanities scholarship », *Literary and Linguistic Computing*, vol. 13, n° 3, 1998, p. 111-117.

Hoover D.L. « Testing Burrows's delta », *Literary and Linguistic Computing,* vol. 19, n° 4, 2004, p. 453-475.

Hoover D.L. « Corpus Stylistics, Stylometry, and the styles of Henry James », *Style*, vol. 41, n° 2, 2007, p. 160-189.

Juola P. « Authorship attribution », *Foundations and Trends in Information Retrieval*, vol. 1, n° 3, 2006.

Labbé D. *Si deux et deux font quatre, Molière n'a pas écrit Dom Juan*, Paris, Max Milo, 2009.

Love H. *Attributing Authorship: An Introduction*, Cambridge University Press, Cambridge, 2002.

Manning C.D., Raghavan P. & Schütze H. *Introduction to Information Retrieval*, Cambridge, Cambridge University Press, 2008.

Manning C.D. & Schütze H. *Foundations of Statistical Natural Language Processing,* Cambridge, The MIT Press, 1999.

Monière D. & Labbé D. « L'influence des plumes de l'ombre sur les discours des politiciens », *Actes JADT*, Besançon, 2006, pp. 687-696

Mosteller F. & Wallace D.L. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, Reading (MA), Addison-Wesley, 1964.

Ng H.T., Goh W.B. & Low K.L. « Feature selection, perceptron learning, and a usability case study for text categorization », *Proceedings ACM-SIGIR*, 1997, p. 67-73.

Peters C., Braschler M., Gonzalo J. & Kluck M. (Eds). *Comparative Evaluation of Multilingual Information Access Systems*. Berlin, Springer-Verlag, LNCS #3237, 2004.

Sebastiani F. « Machine learning in automatic text categorization », *ACM Computing Survey*, vol. 14, n° 1, 2002, p. 1-27.

Sichel H.S. « On a distribution law for word frequencies », *Journal of the American Statistical Association*, vol. 70, n° 351, 1975, p. 542-547.

Stamatatos E. « A survey of modern authorship attribution methods », *Journal American Society for Information Science and Technology*, vol. 60, n° 3, 2009, p. 433-214.

Yang Y. & Pedersen J.O. « A comparative study of feature selection in text categorization », In *Proceedings ICML*, 1997, p. 412-420.

Yang Y. « An evaluation of statistical approaches to text categorization », Information Retrieval, vol. 1, n° 1-2, 1999, p. 69-90.

Yang, Y., & Liu, JX. « A re-examination of text categorization methods », In *Proceedings of the ACM-SIGIR'1999*, p. 42-49

Zhao Y. & Zobel J. « Effective and scalable authorship attribution using function words », *Proceedings of AIRS*, 2005, Berlin, Springer-Verlag, p. 174-189.

Zhao Y. & Zobel J. « Searching with style: Authorship attribution in classic literature », *Proceedings ACSC2007*, 2007, Ballarat, p. 59-68.

Zheng R., Li J., Chen H. & Huang Z. « A framework for authorship identification of online messages: Writing-style features and classification techniques », *Journal of the American Society for Information Science & Technology*, vol. 57, n° 3, 2006, p. 378-393.

9. Annexe

DIA(t_k,c_i)	$\text{Prob}[c_i \mid t_k]$
IMP(t_k,c_i)	$\log_2\left[\text{Prob}[t_k,c_i] \Big/ \text{Prob}[t_k]\cdot\text{Prob}[c_i]\right] = \log_2\left[\text{Prob}[t_k \mid c_i]\right] - \log_2\left[\text{Prob}[t_k]\right]$
OR(t_k,c_i)	$\text{Prob}[t_k \mid c_i]\cdot\left(1-\text{Prob}[t_k \mid -c_i]\right) \Big/ \left(1-\text{Prob}[t_k \mid c_i]\right)\cdot\text{Prob}[t_k \mid -c_i]$
GI(t_k,c_i)	$\sum_{c\in\{c_i,-c_i\}}\ \sum_{t\in\{t_k,-t_k\}} \text{Prob}[t,c]\cdot\log_2\left[\text{Prob}[t,c] \Big/ \text{Prob}[t]\cdot\text{Prob}[c]\right]$
χ^2(t_k,c_i)	$\dfrac{n\cdot\left[\left(\text{Prob}[t_k,c_i]\cdot\text{Prob}[-t_k,-c_i]\right)-\left(\text{Prob}[t_k,-c_i]\cdot\text{Prob}[-t_k,c_i]\right)\right]^2}{\text{Prob}[t_k]\cdot\text{Prob}[-t_k]\cdot\text{Prob}[c_i]\cdot\text{Prob}[-c_i]}$
CC(t_k,c_i)	$\dfrac{\sqrt{n}\cdot\left[\left(\text{Prob}[t_k,c_i]\cdot\text{Prob}[-t_k,-c_i]\right)-\left(\text{Prob}[t_k,-c_i]\cdot\text{Prob}[-t_k,c_i]\right)\right]}{\sqrt{\text{Prob}[t_k]\cdot\text{Prob}[-t_k]\cdot\text{Prob}[c_i]\cdot\text{Prob}[-c_i]}}$
GSS(t_k,c_i)	$\left(\text{Prob}[t_k,c_i]\cdot\text{Prob}[-t_k,-c_i]\right)-\left(\text{Prob}[t_k,-c_i]\cdot\text{Prob}[-t_k,c_i]\right)$

Tableau A.1. *Liste des fonctions utilisées pour la sélection des termes avec leur équation correspondante*

	Estimation	Assoc. pos.	Indép.
DIA(t_k,c_i)	a / (a+b)		
IMP(t_k,c_i)	$\log_2[a\cdot n$ / (a+b)·(a+c)]	>> 0	≈ 0
OR(t_k,c_i)	(a · d) / (c · b)	> 1	≈ 1
GI(t_k,c_i)	$a/n\cdot\log_2[a\cdot n$ / (a+b)(a+c)] + $b/n\cdot\log_2[b\cdot n$ / (a+b)(b+d)] + $c/n\cdot\log_2[c\cdot n$ / (a+c)(c+d)] + $d/n\cdot\log_2[d\cdot n$ / (b+d)(c+d)]	>> 0	≈ 0
χ^2(t_k,c_i)	$n\cdot(a\cdot d - c\cdot b)^2$ / [(a+c)·(b+d)·(a+b)·(c+d)]	>> 1	≈ 0
CC(t_k,c_i)	sqrt(n) · (a·d - c·b) / sqrt[(a+c)·(b+d)·(a+b)·(c+d)]	>> 0	≈ 0
GSS(t_k,c_i)	[(a·d) - (c·d)] /n^2	>> 0	≈ 0

Tableau A.2. *Estimation des fonctions de sélection et les indices permettant de définir une association positive ou l'indépendance*