Détection et regroupement automatique de style d'écriture dans un texte

Jérémy Ferrero*, Alain Simac-Lejeune**

Compilatio
276, rue du Mont-Blanc
74520 Saint-Félix, France
*jeremyf@compilatio.net
**alain@compilatio.net

Résumé. La détection de plagiat extrinsèque devient vite inefficace lorsque l'on n'a pas accès aux documents potentiellement sources du plagiat ou lorsque l'on se confronte à un espace aussi vaste que le Web, ce qui est souvent le cas dans les logiciels anti-plagiat actuels. Dès lors la détection intrinsèque devient nettement plus efficace. Dans cet article, nous traitons justement de la détection automatique d'auteurs qui permet de savoir si un passage d'un texte n'appartient pas au même auteur que le reste du texte et donc en théorie de repérer les passages plagiés d'un document. Nous expliquons notre contribution aux procédures déjà existantes et évaluons les limites de notre approche. L'objectif est de permettre la détection et le regroupement de passages d'un document par auteur.

1 Introduction

La plupart des logiciels anti-plagiat se concentrent sur une détection extrinsèque de plagiat, c'est-à-dire sur le fait de trouver des similitudes entre un document et un corpus de sources probables. Or ce système est inutile si le document ayant été plagié ne se trouve pas dans le corpus fouillé. Néanmoins, il existe un autre type de détection, la détection intrinsèque qui exploite des données extraites de l'intérieur même du document. La détection d'auteurs par étude du style d'écriture du document est la forme de détection de plagiat intrinsèque la plus répandue. Cette approche diverge selon les travaux car elle soulève plusieurs problèmes, allant du découpage du texte de façon pertinente, au choix et à la collecte des données stylistiques à surveiller, en passant par la manière de découper et de classer les différents passages du document par auteur. C'est sur ce dernier point que l'article va essentiellement se concentrer.

2 La détection d'auteur

2.1 La notion de stylométrie

La stylométrie ou l'étude stylométrique d'un texte est une analyse à mi chemin entre une analyse linguistique et statistique. Elle exploite des variables stylométriques, qui sont des

caractéristiques linguistiques du texte, afin d'établir des statistiques sur le document étudié. Effectuer l'analyse stylométrique d'un document consiste à surveiller les variations du style d'écriture du document en surveillant l'évolution des variables stylométriques au sein de celui-ci afin d'en détecter les irrégularités et ainsi pouvoir déterminer si certains passages, appelés phases stylistiques, sortent de la norme par rapport à la majorité du texte.

2.2 État de l'art

Dès le XIX^{e} siècle, Mendenhall (1887) suggère qu'en analysant des caractéristiques internes d'un texte on peut en reconnaître l'auteur. Depuis, les techniques d'études stylométriques de document ont fait d'importantes avancées et de nombreuses recherches (Stein et Eissen, 2007; Layton et al., 2013; Jayapal et Goswami, 2013) appliquent cette découverte à la détection de plagiat. Certaines de ces recherches se concentrent sur l'extraction et la surveillance des données stylométriques les plus pertinentes. Stein et Eissen (2007) ainsi que Zamani et al. (2014) surveillent les proportions d'utilisation des classes lexicales au sein de segments afin de discerner lesquelles sont les plus porteurs du style de l'auteur. Oberreuter et Velásquez (2013) privilégient, quant à eux, la fréquence des termes comme donnée stylométrique à surveiller.

Parallèlement à cela, des recherches d'apprentissage et classification (Layton et al., 2013) ont vu le jour. Elles caractérisent le style d'un auteur par un modèle de langage n-grammes, ainsi ils entraînent leur module sur un corpus d'apprentissage d'auteurs et appliquent le modèle en résultant sur des passages tests afin d'en calculer l'auteur le plus probable.

3 Notre approche

Une segmentation pseudo-sémantique soutenue par le travail de Zechner et al. (2009) a été utilisée ainsi que l'idée de van Halteren (2004) qui introduit la notion de surveillance de plusieurs variables stylométriques. Notre contribution réside essentiellement dans le découpage et le regroupement automatique des phases stylistiques par auteur au moyen d'une implémentation spécifique à notre problème d'un Mean Shift (Cheng, 1995), un algorithme multidimensionnel des k-moyennes non paramétrique.

3.1 Segmentation

Dans un premier temps, l'idée est de segmenter le document. Il est important que chaque segment conserve un sens afin d'être autonome et donc d'être potentiellement écrit par une personne différente. Une segmentation en unité de sens est donc à privilégier. S'appuyant sur le travail de Zechner et al. (2009), c'est une segmentation pseudo sémantique qui a été retenue : un découpage par phrase d'une taille minimale (en mots). Le seuil a été fixé à 15 mots, taille moyenne des phrases dans la langue française.

3.2 Extraction de la stylométrie

La seconde étape du processus consiste à extraire la stylométrie de chaque segment. Pour ce faire, il faut au préalable détecter la langue de chaque segment au moyen d'un module im-

plémentant la technique de catégorisation de texte à base de n-grammes de Cavnar et Trenkle (1994). Ensuite, l'étiqueteur morphosyntaxique TreeTagger (Schmid, 1994) a été utilisé afin de déterminer la classe lexicale de chaque unité lexicale du texte. De cette façon l'étiquetage morphosyntaxique s'effectue en fonction de la langue du segment et notre procédé est indépendant de la langue du texte. Enfin, un calcul du pourcentage d'utilisation de chaque classe lexicale extraite est effectué. Les variables stylométriques utilisées sont les taux d'utilisation par segment des noms propres et communs, des verbes, adjectifs, adverbes, pronoms, prépositions, virgules et nombres. La taille moyenne des mots et la taille moyenne des phrases sont également calculées et ajoutées aux données stylométriques.

3.3 Construction des courbes

Une fois la segmentation et les calculs stylométriques opérés, on obtient donc plusieurs valeurs par segment (i.e. une valeur par variable stylométrique). Une suite de valeurs brutes sans cohérence n'étant pas exploitable, on représente la stylométrie du document sous la forme de courbes, avec en abscisse, la position des segments (la ligne de vie du document) et en ordonnée, les valeurs des variables stylométriques observées. Ceci a pour avantage, en plus de permettre une représentation visuelle, de faciliter la comparaison et la manipulation des valeurs entre elles, les algorithmes de manipulation de courbes étant courant.

Il est possible que le style d'un même auteur varie énormément au fil d'un même texte. La fatigue ou la maturité lors de longs écrits peuvent entraîner du bruit ou des variations brusques. On convient alors qu'un lissage est nécessaire. C'est le lissage par la moyenne glissante sans pondération (Chou, 1975) qui a été utilisé dans cet article.

3.4 Regroupement

Il reste à déterminer les phases stylistiques de façon automatique. Un algorithme d'apprentissage automatique non supervisé (i.e. sans intervention humaine) est idéal dans ce cas de figure qui s'apparente au clustering car il faut déterminer à quel auteur (i.e. à quel cluster) chaque donnée s'apparente. Sachant que le nombre d'auteurs et donc de clusters n'est pas connu à l'avance, c'est le Mean Shift multidimensionnel (Cheng, 1995) qui se dégage. En effet, cet algorithme permet de clustériser un ensemble de points sans connaître à l'avance le nombre k de clusters. L'idée dans notre cas est de déterminer le nombre k à partir d'un seuil. On définit alors empiriquement un nombre k de départ assez grand, admettons 10 et un seuil, entre 2% et 15% en fonction de la moyenne de la variable stylométrique observée (seuil adaptatif). Tant qu'il existe deux clusters voisins avec une différence de stylométrie inférieure au seuil, on relance un KMeans avec $k = k - 1$. Une fois toutes nos phases identifiées et k définitif, s'il existe deux clusters (non voisins cette fois-ci étant donné que les voisins ont déjà été réunifiés) avec une différence de stylométrie inférieure au seuil, on en déduit qu'ils sont du même auteur.

On prend en considération plusieurs variables stylométriques en même temps, tout comme le fait van Halteren (2004). L'idée est de surveiller plusieurs variables stylométriques afin qu'elles se « complètent » mutuellement. On augmente ainsi le taux de certitude de l'existence d'une zone par le fait qu'une zone est définie comme telle si la majorité des courbes fléchissent de telle façon à la dessiner. De plus, la zone de flexion retenue est maintenant désignée par la moyenne des zones de flexion de toutes les courbes surveillées, ceci réduisant considérablement l'erreur d'approximation et rendant plus sûr notre prise de décision. Pour exemple, sur la

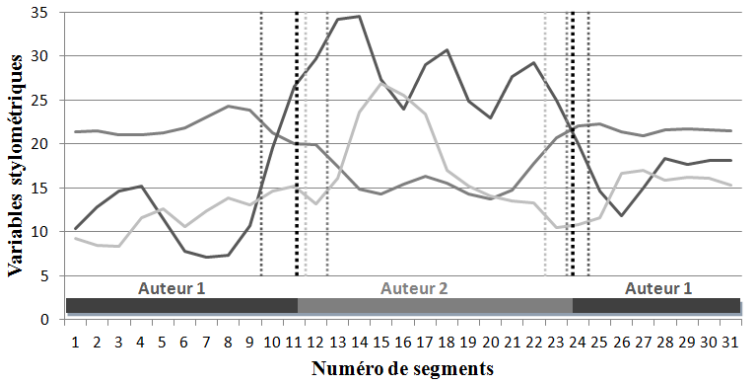

FIG. 1 – *Mean Shift sur plusieurs variables stylométrique.*

figure 1 chaque zone de flexion des courbes est représentée par une ligne verticale pointillée de la même couleur que la courbe dont elle dépend. Les lignes pointillées noires plus épaisses représentent les découpages retenus (les moyennes des trois flexions des trois courbes).

Pour faciliter l'observation des écarts et des flexions, les courbes sur cette figure ont été normalisées (mises à la même échelle), leurs valeurs stylométriques sont donc faussées.

4 Évaluation et tests

4.1 La base de tests et protocole

La base de tests est composée de 500 textes contenant en moyenne 7 000 mots. Les textes sont constitués d'un (l'intégralité du texte) à cinq passages, chaque passage étant potentiellement écrit par un auteur différent. Un texte peut contenir plusieurs passages écrits par un même auteur. On recense en totalité dans la base, une dizaine d'auteurs différents. La langue prédominante au sein des textes est le français, cependant pour tester l'adaptabilité et le plurilinguisme du système de nombreux passages sont en anglais ou en italien. Afin de tester correctement la procédure évaluée, des passages traitant du même sujet et donc employant le même vocabulaire ont été utilisés dans le but de tromper la stylométrie extraite. De plus, l'intégralité des textes est annotée, de telle sorte à savoir précisément de quel mot à quel mot les textes sont écrits par un auteur ou par un autre.

4.2 Résultats

Notre procédure présente une précision de 0.89 et un rappel de 0.34. Il est néanmoins important d'étudier plus en détails les limites de cette procédure et de nuancer un rappel si faible. La figure 2 est un diagramme à bulles représentant les performances du découpage stylométrique. L'axe des abscisses représente la taille en segment de la phase stylistique concernée et celui des ordonnées l'écart moyen de la variable stylométrique observé entre cette phase et ses voisines, son unité est notée *us* pour unité stylométrique. Les bulles représentent les différents

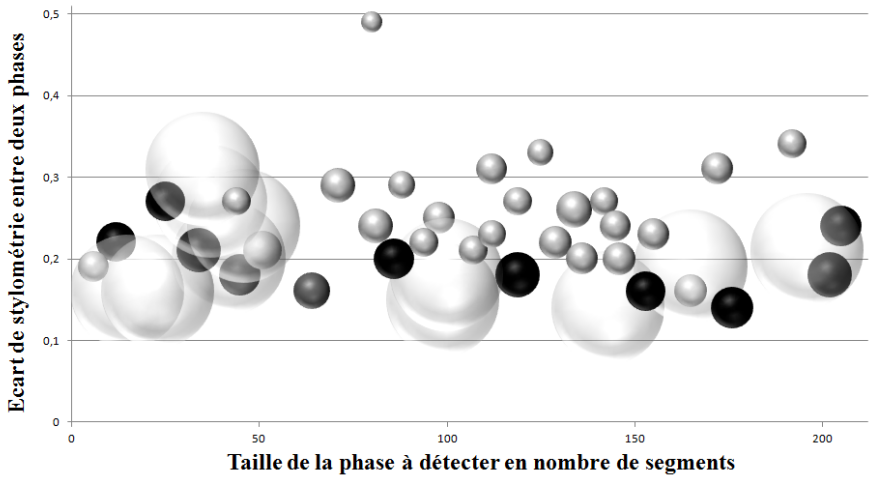

FIG. 2 – *Performance du Mean Shift en fonction de la taille des phases et de l'écart des variables.*

tests effectués et leur taille, le taux d'approximation qu'il en a découlé. Les grandes bulles blanches transparentes sont des erreurs de 100%, là où la segmentation a été complètement fausse et le Mean Shift a décroché, retournant un seul cluster. Les bulles en noirs sont les tests où la marge d'erreur a été supérieure à 10% et donc où les résultats restent peu exploitables. Dans un souci de clarté, seul les tests pertinents et non redondants ont été conservés lors de l'affichage dans la figure 2. Une zone de confort où le taux d'erreur d'approximation du découpage reste constamment inférieur à 8% se dégage. Cette zone se situe lorsqu'on essaie de détecter une zone comprise entre 40 et 190 segments et lorsque l'écart stylométrique entre deux phases est supérieur à 0.20 *us*. En considérant seulement ces cas, le rappel augmente à 0.62 et la précision passe à 0.70.

5 Conclusions

Notre approche montre des résultats exploitables lorsque les phases stylométriques à identifier ne sont pas trop importantes (n'excèdent pas 190 segments soit environ 4000 mots) et lorsque la différence de stylométrie est suffisamment grande (supérieur à 0.20 *us*). En revanche dans tous autres cas, les limites de notre approche se font ressentir. Pour palier ces problèmes, un seuil adaptatif pourra être défini en fonction du type de variable stylométrique surveillée. De plus, avec du recul, nous convenons qu'un Mean Shift n'était sans doute pas la meilleure option de clustering. Dans la suite de nos travaux nous implémenterons d'autres classifieurs (hiérarchique, DBSCAN, etc.).

Pour conclure, bien que perfectible, cette approche permet de détecter différents styles d'écriture au sein d'un même texte et notre contribution malgré ses limites permet bien de regrouper automatiquement les phases stylistiques par auteur.

Références

Cavnar, W. B. et J. M. Trenkle (1994). N-Gram-Based Text Categorization. In *Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175.

Cheng, Y. (1995). Mean Shift, Mode Seeking, and Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17*(8), 790–799.

Chou, Y. (1975). *Statistical analysis : with business and economic applications*. Holt, Rinehart and Winston Quantitative Methods Series. Holt International.

Jayapal, A. K. et B. Goswami (2013). Vector space model and overlap metric for author identification. In *Notebook for PAN at CLEF 2013*.

Layton, R., P. Watters, et R. Dazeley (2013). Local n-grams for author identification. In *Notebook for PAN at CLEF 2013*.

Mendenhall, T. C. (1887). The Characteristic Curves of Composition. In *Science IX*, Volume 102, pp. 237–249.

Oberreuter, G. et J. D. Velásquez (2013). Text mining applied to plagiarism detection : The use of words for detecting deviations in the writing style. In *Expert Systems with Applications*, Volume 40, pp. 3756–3763.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Stein, B. et S. M. Z. Eissen (2007). Intrinsic Plagiarism Analysis with Meta Learning. In B. Stein, M. Koppel, et E. Stamatatos (Eds.), *PAN*, Volume 276 of *CEUR Workshop Proceedings*.

van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In *Association for Computational Linguistics*, Volume 42, pp. 199–206.

Zamani, H., H. Nasr, P. Babaie, S. Abnar, M. Dehghani, et A. Shakery (2014). Authorship identification using dynamic selection of features from probabilistic feature set. In *CLEF 2014*, pp. 128–140.

Zechner, M., M. Muhr, R. Kern, et M. Granitzer (2009). External and Intrinsic Plagiarism Detection Using Vector Space Models. In Stein, Rosso, Stamatatos, Koppel, et Agirre (Eds.), *PAN'09*, pp. 47–55.

Summary

Extrinsic plagiarism detection quickly becomes ineffective when you do not have access to potentially sources documents of plagiarism or when the search space is large like the Web, which is often the case with current anti-plagiarism software. Therefore the intrinsic detection becomes much more effective. In this paper, the automatic authorship detection is exactly presented. It allows to know if a text's part does not belong to the same author as the rest of the text and so in theory to identify plagiarized passages of a document. We explain our contribution to the existing procedures and assess the limitations of our approach. The goal is to enable the detection and clustering of passages in a document by author.