Analyse comparative de méthodes d'apprentissage pour la catégorisation de textes selon leur langue de rédaction

Baptiste Bohet*, Nicole Vincent**

* Université Sorbonne Nouvelle
baptiste.bohet@sorbonne-nouvelle.fr,
** Université Paris Cité, LIPADE, F-75006 Paris, France
nicole.vincent@u-paris.fr

Résumé. L'objectif de cette étude est double. Il s'agit, d'une part, de catégoriser des textes romanesques en français pour permettre à un utilisateur de déterminer s'ils sont originaux ou traduits, c'est-à-dire nativement rédigés en français ou non. D'autre part, de procéder à une analyse comparative et d'optimiser les méthodes choisies pour obtenir ce résultat. Les données textuelles considérées ici sont volumineuses, variées thématiquement et stylistiquement. Les quatre méthodes mises en œuvre – qui prennent en compte aussi bien les caractéristiques fréquentielles, que lexicales, syntaxiques ou sémantiques – reposent sur un apprentissage automatique. L'analyse comparative des approches porte sur l'espace de représentation des données, le paramétrage, les taux de classifications (par classes et global) et l'explicabilité.

1 Introduction

L'objectif de cette étude est d'être capable d'identifier si un texte fictionnel, présenté en français est original, c'est-à-dire s'il a été nativement écrit en français (nous le désignerons par *TO* pour Texte Original), ou s'il est le fruit d'une traduction (*TT* pour Texte Traduit). Le système réalisé, à terme implémenté sur un site web, doit permettre à un utilisateur d'entrer un texte, ou une portion de texte, et d'obtenir la nature de celui-ci (*TT* ou *TO*). Précisons qu'il ne sera pas question ici de traduction automatique, car aucun éditeur n'en publie pour la littérature romanesque qui correspond à notre corpus. Tous les romans publiés actuellement en français sont traduits par des humains.

L'idée qui préside à ce projet, est que certains lecteurs experts, sans informations liées au paratexte éditorial (ici le nom de l'auteur et la langue originale), sont parfois capables d'identifier si le texte qu'ils lisent est originairement écrit en français ou s'il s'agit d'une traduction. S'ils sont capables de cette identification, ils ne sont pas pour autant en mesure d'expliquer les éléments qui motivent leur intuition. Il nous a donc semblé intéressant de voir si, et comment, différentes techniques d'apprentissage machine pouvaient réaliser une telle discrimination.

Cela est d'autant plus intéressant, que le résultat de cette étude devrait notamment permettre, en l'associant à d'autres, de faire des recherches en paternité, d'identifier des textes apocryphes, ou encore de lutter contre le plagiat.

Pour mener à bien cette étude, nous avons comparé quatre méthodes dans la double perspective de retenir la plus efficiente et d'identifier les éléments permettant la classification :

— *FFA* (*Few Feature Approach*), exploite des caractéristiques lexicales ;
— *FA* (*Frequency Approach*), considère les fréquences lexicales ;
— *RFA* (*Relative Frequency Approach*) basée sur *TF-IDF* (*Term Frequency-Inverse Document Frequency*), s'appuie sur les fréquences relatives ;
— *SA* (*Semantic Approach*) basée sur le modèle *BERT* (*Bidirectional Encoder Representations from Transformers*) prend en compte la proximité lexicale.

Après un panorama des études portant sur des thématiques proches, nous présenterons les choix qui ont présidé à la constitution de notre corpus. Nous exposerons ensuite les spécificités des quatre approches, puis nous comparerons les résultats obtenus en fonction de la méthode utilisée. Enfin, la conclusion nous permettra d'envisager les perspectives offertes par l'étude.

Études connexes. L'informatique, dès ses débuts, a contribué à l'analyse et à la classification de textes écrits en langage naturel. On trouve dans (Gasparetto et al., 2022) un état des lieux très complet sur ce sujet. Certaines classifications s'intéressent au contenu (Shen et al., 2018), d'autres, à des aspects plus stylistiques et sociolinguistiques (Marteau et Vincent, 2006), ou plus formels, par exemple, la langue utilisée (Sundermeyer et al., 2012).

On trouve aussi des études qui analysent les particularités entre les traductions humaines et celles automatiques (rappelons que, dans notre cas, les textes traduits le sont par des humains). Dans (Popovic, 2020) les caractéristiques utilisées portent sur la longueur des phrases et des mots, la richesse du vocabulaire, les statistiques liées à un étiquetage morphosyntaxique du texte. D'autres études se fondent encore sur des n-grams, ou prennent en compte la présence/absence de termes d'une liste prédéfinie comme dans (Aharoni et al., 2014). Dans (Rabinovich et al., 2016), sont utilisés des tri-words, des fréquences de mots en fonction de leur position dans la phrase et celles des mots fonctionnels. Notons que la plupart de ces études ont été menées sur des documents administratifs et non des textes fictionnels.

Plus récemment, l'analyse des textes a eu largement recours aux modèles *BERT*. On notera, par exemple, l'étude sur les émotions exprimées dans des tweets (Chiorrini et al., 2021).

2 Description du corpus

La première étape pour mener à bien notre étude a consisté à constituer un corpus d'apprentissage. Nous avons sélectionné 2 000 textes : 1 000 *TO* / 1 000 *TT*. Cela représente environ 1,2 Go de textes bruts (txt), soit 184 422 466 mots, soit 1 067 646 686 caractères.

Le corpus est strictement constitué de romans, pour ceux traduits, ils proviennent exclusivement de la langue anglaise pour éviter une multitude de particularités idiomatiques. Par ailleurs, pour éviter les biais chronologiques liés à l'évolution de la langue, les textes sont tous contemporains (postérieurs à 1980). Précisons enfin que les textes choisis proviennent de tous les genres littéraires, qu'ils n'ont pas de thématiques communes. Le corpus test répond aux mêmes exigences que celui pour l'apprentissage, il est composé de 200 textes : 100 *TO* / 100 *TT*. Pour éviter les biais cognitifs, le corpus test est strictement composé d'auteurs distincts.

3 Méthodologie

Ici, chacune des méthodes transpose le problème de classification de textes en un problème de classification de vecteurs.

Méthode *FFA* : Cette première méthode est basée sur trois caractéristiques lexicales : La *Richesse Lexicale* (*RL*) ; La *BiVersité* (*BV*) ; La *Proportion des Hapax* (*PH*).

La *Richesse Lexicale* (*RL*) est le rapport entre le nombre de mots différents (ce que les spécialistes de lexicométrie appellent une “forme”) donc nbf et le nombre de mots total du texte (les "occurrences") donc nbo (Bernard et Bohet, 2017) : $RL = nbf/nbo$.

Le néologisme *BiVersité* (*BV*), est un concept innovant que nous avons créé, il représente la proportion des formes dont la fréquence absolue est supérieure ou égale à deux. C'est-à-dire, le rapport entre le nombre de formes qui ne sont pas des hapax (des formes n'ayant qu'une occurrence dans le texte) et le nombre total de formes : $BV = (nbf - nbh)/nbf$.

Enfin, nous entendons par *Proportion des Hapax* (*PH*), le rapport entre le nombre d'hapax nbh et le nombre d'occurrences dans le texte nbo : $PH = nbh/nbo$

Utilisées individuellement pour construire un classifieur SVM à noyau RBF, chaque caractéristique *RL*, *BV* et *PH* conduit à un taux de reconnaissance respectivement de 71,2%, 74,1% et 72,3%. La variable *BV* est la plus discriminante, mais l'exploitation de l'association des trois caractéristiques améliore les résultats pour atteindre un taux de 75,3%.

Méthode *FA* : Cette méthode, est basée sur l'étude statistique des fréquences. Dans l'approche précédente, nous avons calculé des caractéristiques intrinsèques à chaque échantillon, alors qu'ici chaque échantillon est représenté dans un espace vectoriel à n dimensions où n est le nombre de formes du lexique de l'ensemble du corpus d'apprentissage.

La méthode est organisée comme suit. À partir de l'index de l'ensemble du corpus d'apprentissage, à chacune des formes est associée une dimension de l'espace de représentation. Chaque échantillon est donc représenté dans cet espace par un vecteur V qui prend en compte chacune des formes de l'échantillon. Chaque composante V_i, où i varie de 1 à n, du vecteur V est associée à une forme du corpus et contient la fréquence de celle-ci dans l'échantillon étudié. Ainsi, si la forme f d'indice f_i est présente dans le corpus, mais ne l'est pas dans l'échantillon, on aura $V_{f_i} = 0$. Pour un échantillon de nbo occurrences et nbf formes, le vecteur V comporte nbf composantes non nulles. Pour une forme f présente x fois dans l'échantillon, on a : $V_{f_i} = x/nbo$. Notons que le classifieur utilisé ici est un classifieur bayésien multinomial.

Méthode *RFA* : Cette méthode basée sur la méthode de pondération *TF-IDF* prend en compte les fréquences relatives par rapport à l'ensemble du corpus. La fréquence d'apparition d'une forme n'indique pas sa spécificité par rapport à un document donné ou à un ensemble de documents. Or, une forme commune à de nombreux documents devrait être moins significative qu'une forme commune à peu d'entre eux. La méthode pondérée *TF-IDF* permet de corriger ce biais. *Term Frequency* (*TF*) correspond au nombre d'occurrences d'une forme dans un texte, sa pondération locale, alors que le *Inverted Document Frequency* (*IDF*) désigne la valeur inverse du nombre de documents dans lesquels la forme est présente, autrement dit sa pondération globale. Ainsi, la combinaison *TF-IDF* met en exergue le nombre d'occurrences de la forme dans le document par rapport à sa distribution dans l'ensemble du corpus. Ceci dans la perspective

d'évaluer sa pertinence (sa surreprésentation ou, au contraire, sa rareté relative). Un classifieur multinomial bayésien est utilisé, la pondération de la forme f dans un document d est calculée par [$TF - IDF_{f,d} = TF_{f,d}.(log(\frac{|D|}{DF_f}) + 1)$] avec |D| le nombre de documents, $TF_{f,d}$ la fréquence de la forme f dans le document d et DF_f le nombre de documents comprenant la forme f.

Méthode *SA* : Les méthodes précédentes reposent sur du *shallow learning*, il nous a semblé pertinent de pouvoir les comparer à une méthode relevant du *deep learning*. Dans la mesure où nous travaillons sur des données textuelles, l'architecture *BERT* (Devlin et al., 2019) paraît être le choix le plus judicieux. Notre corpus est en français, nous utilisons donc CamemBERT.

Pour mémoire, *BERT* est un modèle pré-entraîné capable de résoudre plusieurs problématiques de Traitement de Langage Naturel, basé sur le concept de *mécanisme d'attention*. Le réseau agit de façon sélective, pour ne se concentrer que sur quelques éléments pertinents, tout en ignorant les autres. Il comprend deux mécanismes distincts : un encodeur qui lit l'entrée de texte et le transforme en vecteur et un décodeur qui produit une prédiction pour la tâche.

Dans notre projet, seul le premier mécanisme, l'encodeur, est nécessaire, il est suivi d'une étape de classification. À une séquence de mots en entrée, ou pour être plus précis, de tokens (*BERT* tokenise les textes au préalable) de longueur l, est attribuée une séquence de vecteurs de longueur l, dans laquelle chaque vecteur correspond à un élément d'entrée. Ce passage à une représentation vectorielle est réalisé grâce à un apprentissage sur un vaste ensemble de textes de sorte que les vecteurs traduisent une proximité sémantique entre éléments. Le classificateur utilisé est un *fully connected*, ici *CamembertForSequenceClassification*.

Principe de décision : Une fois les apprentissages réalisés, le texte à tester est découpé en échantillons. C'est un vote majoritaire qui intervient alors pour la prise de décision.

4 Résultats expérimentaux

Les apprentissages sont réalisés par une validation croisée à 5 folds (5-fold CV), aléatoires et les résultats présentés ici correspondent à la valeur moyenne à l'issue de 10 apprentissages.

Comparaison des résultats des quatre méthodes : Les résultats obtenus sont les suivants :

	FFA	FA	RFA	SA
TR de l'apprentissage	74,6%	98,1%	98,9%	**99,5%**
TR du test utilisateur	75,5%	94,0%	95,0%	97,5%

TAB. 1 – Taux de reconnaissance (*TR*) des quatre méthodes.

Les résultats obtenus sont remarquablement élevés, alors que le problème semblait extrêmement difficile. Il apparaît ainsi que les traductions, même si elles sont réalisées par des traducteurs humains chevronnés, ont des caractéristiques spécifiques qui permettent, avec les outils que nous avons mis en œuvre, de discriminer efficacement des textes nativement écrits

en français. En effet, si *FFA* a des résultats mitigés (environ 75%), les trois autres méthodes ont, elles, des résultats supérieurs à 90%. Notons la performance admirable de *SA* qui avec le système d'apprentissage a un taux de reconnaissance de 99,5%, et avec notre échantillon de test utilisateur a un taux de reconnaissance de 97,5%. L'objectif de notre étude est atteint bien au-delà de nos attentes : il est possible de classifier des textes selon qu'ils sont nativement écrits en français ou le fruit d'une traduction effectuée par un traducteur humain.

La question demeure pourtant de savoir quelles caractéristiques permettent une catégorisation si efficace, autrement dit, d'interroger l'explicabilité des résultats obtenus. Dans cette perspective, il est intéressant de constater que l'explicabilité est inversement proportionnelle à la qualité des résultats. De fait, *FFA* qui offre des pistes de réflexion relativement facilement interprétables ne propose qu'un taux de reconnaissance modeste.

Ajoutons que, puisque les méthodes choisies reposent sur des apprentissages établis à partir de caractéristiques différentes, force est de constater que les éléments permettant la classification sont multiples. Il n'y a pas un sésame donnant la clé de l'énigme, mais une série d'éléments discriminants : les différences sont multifactorielles. Non seulement les textes sont aisément catégorisables, mais qui plus est, ils le sont en utilisant plusieurs critères. La fréquence des formes, la sous ou surreprésentation de certaines d'entre elles, la proximité cooccurrencielle sont autant d'indices efficaces pour cette catégorisation.

Influence de la taille et nombre des échantillons d'apprentissage : Pour montrer que la taille des échantillons d'apprentissage a une influence sur la qualité des résultats, nous avons expérimenté différents scénarios, dans la perspective d'optimiser les résultats des classifieurs. La Figure 1 montre que *FFA* gagne en efficacité quand la taille des échantillons augmente. Ceci est aussi vrai pour les trois autres méthodes, mais dans des proportions moindres. Ceci s'explique probablement en partie, car même avec des échantillons de taille réduite, les résultats sont déjà très élevés et qu'ils ne peuvent donc qu'augmenter faiblement.

FIG. 1 – Taux de reconnaissance en fonction de la taille des échantillons d'apprentissage (en signes pour *FFA, FA* et *RFA*, en tokens pour SA).

Influence de l'augmentation des données d'apprentissage : Les études menées ici utilisent intégralement les textes, la quantité de données est donc fixe. Pour améliorer nos résultats, nous avons réalisé une augmentation des données en considérant des portions chevauchantes de textes (non synthétiques), multipliant ainsi par deux ou quatre le nombre d'échantillons à taille constante. Nous assurons ainsi, non seulement une augmentation du nombre d'échantillons, mais aussi une amélioration de la diversité des représentations des échantillons.

Le Tableau 2, relatif à *FFA, FA* et *RFA* montre que les résultats s'améliorent lorsque les échantillons de 200 000 signes sont plus nombreux. L'augmentation conduit aussi à une plus

		FFA	FA	RFA
Échantillon 200 000 signes	*TR*	73,7%	96,8%	97,5%
Sans augmentation	écart-type	1,25	0,39	0,81
Échantillon 200 000 signes	*TR*	74,2%	98,0%	98,1%
Augmentation $\times 2$	écart-type	0,90	0,20	0,43
Échantillon 200 000 signes	*TR*	74,6%	98,1%	98,9%
Augmentation $\times 4$	écart-type	0,39	0,19	0,11

TAB. 2 – Taux de reconnaissance (*TR*) et écart-type en fonction de l'augmentation des données.

grande fiabilité des résultats, comme en attestent les écarts-types qui diminuent de manière significative. Malgré l'augmentation, le taux de reconnaissance de *FFA* reste lui relativement faible, ce qui laisse à penser que les caractéristiques choisies ne sont pas suffisantes pour la tâche à accomplir. Si *FA* paraît avoir atteint un nombre d'échantillons suffisant pour l'apprentissage, *RFA*, semble, elle, pouvoir encore être améliorée en augmentant la taille de l'ensemble d'apprentissage. Soulignons enfin que pour les trois méthodes, il est plus significatif de regarder l'évolution des taux d'erreur. De fait, si l'on prend l'exemple de *RFA*, son taux d'erreur sans augmentation de 2,5% passe à 1,1% avec une augmentation de $\times 4$, soit une amélioration de 60%.

Influence de la taille des échantillons pour l'analyse lors du test utilisateur : La taille des échantillons, s'agissant du test utilisateur, influe elle aussi sur les résultats. Comme on peut l'observer sur la Figure 2, l'influence est bien différente selon les méthodes. De fait, *FFA* ne donne de résultats probants qu'avec des échantillons supérieurs à 50 000 signes. A contrario, *FA, RFA et SA* ne s'améliorent que relativement peu en fonction de la taille des échantillons. Il est par ailleurs remarquable de constater que *FA, RFA* et *SA* sont capables de taux de reconnaissance supérieurs à 90% avec des échantillons de test très réduits : 1 000 signes ou 128 tokens ne représentent en effet qu'à peine une page imprimée.

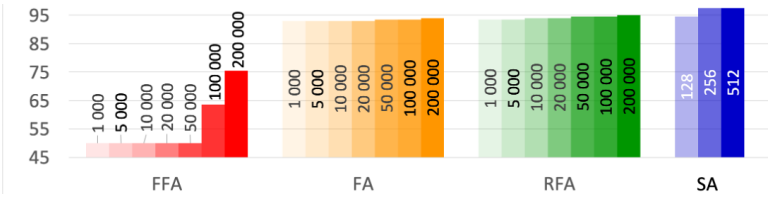

FIG. 2 – Taux de reconnaissance en fonction de la taille de l'échantillon lors du test utilisateur (en signes pour *FFA, FA* et *RFA*, en tokens pour *SA*).

Comparaison entre résultats traduits et originaux : L'un des éléments novateurs et riches en enseignements de cette étude est d'observer sur les matrices de confusion, Figure 3, que le taux de reconnaissance varie selon la méthode, mais aussi en fonction des classes.

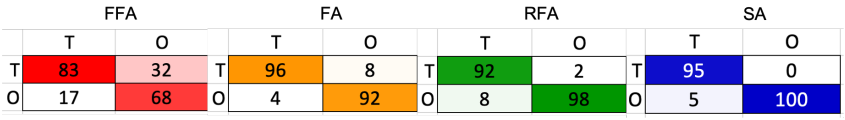

FIG. 3 – Matrices de confusion relatives aux 200 échantillons du test utilisateur.

En effet, si *FFA* et *FA* reconnaissent mieux les *TT*, *RFA* et *SA* sont, elles, meilleures pour la reconnaissance des *TO*. Nous ne pouvons ici que proposer des hypothèses, car il faudrait poursuivre davantage les analyses pour être catégoriques, mais il semble donc que les systèmes purement formels soient plus efficaces avec les *TT* alors que ceux ayant une approche plus sémantique identifient mieux les *TO*. *RFA* (basée sur *TF-IDF*), reconnaît 98% des originaux comme tels ; *SA* (basée sur *BERT*) catégorise 100% des originaux comme tels. L'une et l'autre sont moins performantes avec les *TT* puisque respectivement, elles considèrent 8 (*RFA*) et 5 (*SA*) *TT* comme des *TO*. Une hypothèse serait donc que l'identification des *TT* est plus liée au comptage des formes, alors que celle des *TO* serait plus attachée à la proximité des formes entre elles.

5 Conclusion

Toutes les méthodes proposées sont riches en enseignements. S'il n'est, pour l'instant, pas possible d'être définitif sur les conclusions à en tirer, notre analyse fait néanmoins apparaître plusieurs pistes.

Première découverte, la catégorisation est multifactorielle. La syntaxe, les caractéristiques lexicométriques, les fréquences, les proximités sémantiques sont autant d'indices pour la discrimination. Ceci explique sûrement en partie pourquoi le lecteur, fut-il expert, n'est pas en mesure d'expliquer son intuition.

Ce qui précède ne doit cependant pas oblitérer notre démarche, dans la mesure où les résultats sont extrêmement satisfaisants. Si le *deep learning* avec la méthode *SA* offre la meilleure catégorisation avec près de 97,5% d'élucidation, *FA* et *RFA* ne sont pas en reste avec environ 95%. Ces résultats sont d'autant plus impressionnants que, rappelons-le, les données textuelles traitées sont extrêmement variées (nous avons choisi de travailler sur tous types de romans sans choisir aucun genre qui aurait, à n'en pas douter, facilité l'analyse).

Autre apport de notre travail, l'efficacité des méthodes employées avec de petits échantillons. Il est en effet impressionnant de constater que si l'apprentissage nécessite des corpus conséquents, la classification requiert peu de texte. Notre chapitre sur l'influence du paramétrage montre que la taille des morceaux peut être extrêmement réduite (moins d'une page) sans entamer la qualité des résultats. En l'occurrence, cela fonctionne donc identiquement de la lecture humaine qui ne nécessite pas beaucoup de pages pour identifier la nature linguistique du texte.

Enfin, cette étude ouvre la voie à de nombreuses pistes de recherche que nous avons d'ores et déjà commencé à explorer. Sur le même modèle, en modifiant éventuellement les paramètres, il semble en effet possible d'étudier d'autres questions portant par exemple sur le genre du texte, sa datation, ou plus particulièrement sur l'auteur, son sexe, son âge, ses origines, etc. Cette étude ne représente donc qu'une première étape d'un plus vaste projet.

Remerciements Qu'il nous soit permis de remercier ici Neguin Navidi et Luca Bresolin pour leur travail préparatoire.

Références

Aharoni, R., M. Koppel, et Y. Goldberg (2014). Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, pp. 289–295.

Bernard, M. et B. Bohet (2017). *Littérométrie : outils numériques pour l'analyse des textes littéraires*. Presses Sorbonne nouvelle.

Chiorrini, A., C. Diamantini, A. Mircoli, et D. Potena (2021). Emotion and sentiment analysis of tweets using bert. In *EDBT/ICDT Workshops*.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference on Human Language Technologies*, Minneapolis, Minnesota, pp. 4171–4186.

Gasparetto, A., M. Marcuzzo, A. Zangari, et A. Albarelli (2022). A survey on text classification algorithms : From text to predictions. *Information 13*(2).

Marteau, H. et N. Vincent (2006). Un automate pour évaluer la nature des textes. *Revue des Nouvelles Technologies de l'Information EGC 2006, RNTI-E-6*, 259–270.

Popovic, M. (2020). On the differences between human translations. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal, pp. 365–374. European Association for Machine Translation.

Rabinovich, E., S. Nisioi, N. Ordan, et S. Wintner (2016). On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 1870–1881.

Shen, C., C. Sun, J. Wang, Y. Kang, S. Li, X. Liu, L. Si, M. Zhang, et G. Zhou (2018). Sentiment classification towards question-answering with hierarchical matching network. In *Proceedings of the Conference on Empirical Methods in NLP*, Brussels, pp. 3654–3663.

Sundermeyer, M., R. Schlüter, et H. Ney (2012). LSTM neural networks for language modeling. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pp. 194–197. ISCA.

Summary

The objective of the work is twofold. On the one hand, the aim is to categorize french novels to make it possible for a user to determine whether they are original or translated, that is to say in the original language of the author or not. On the other hand, to compare and optimize the elaborated methods to achieve this goal. Here, the textual data we consider are voluminous and present variety in the themes and styles. The four implemented approaches – taking into account frequency, lexical, syntactic or semantic characteristics – rely on machine learning. The approach comparison considers the representation space as well as the parametrisation of the methods, the recognition rates (by classes or global) or the explainability.