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.