La contribution des outils de Machine Learning et de Deep Learning à l'analyse de textes historique : l'intérêt particulier pour la sémantique historique

Yves Noblet

Année académique 2022/2023

Résumé

La sémantique quantitative constitue un domaine de recherche utile pour la recherche historique.

1 Introduction

Ce travail a pour but de voir comment l'on peut saisir le sens des mots de manière quantitative, autrement dit faire de la sémantique quantitative à l'aide d'outils informatiques. Le but est également de montrer quelques exemples du genre de travail que l'on peut réaliser dans ce domaine, non seulement sur des corpus de textes en langues modernes, mais aussi en langues anciennes. Comme nous allons le voir, les études de sémantique quantitative ne sont pas sans intérêt. Le sens des mots évoluant dans le temps, cela implique par exemple, qu'un historien effectuant une recherche sur un sujet donné pourra ou non incorporer tel ou tel texte comportant les mots clés nécessaires à sa recherche selon si leur sémantique correspond bien à son sujet. La sémantique quantitative apporte donc un certain soutien au chercheur en histoire [5]. Mais l'intérêt des études sémantiques ne s'arrête pas là, en connaissant le sens des mots, on peut effectuer des tâches de classification de textes [5]. On peut également voir quels sont les termes associés à certains mots, ou encore saisir la similarité entre plusieurs termes. On peut voir à quel genre de textes certains mots sont le plus associés [6]. Les possibilités d'études à l'aide d'outils de sémantique quantitative sont donc assez nombreuses et variées.

Afin d'effectuer de la sémantique quantitative, il faut faire appel à des systèmes de Machine Learning, ou encore de Deep Learning, qui n'est rien de moins qu'un sous-ensemble du Machine Learning. Les systèmes de Machine Learning relèvent de l'intelligence artificielle. Avant de poursuivre, il convient de consacrer

une section à la définition de ces notions d'intelligence artificielle, de Machine Learning et de Deep Learning. Une seconde section montre des exemples d'application de l'analyse sémantique dans le domaine de la recherche historique. En premier lieu, quelques exemples d'application sur des textes en langues anciennes et en second lieu sur des textes dans des langues plus modernes.

2 Définitions

2.1 L'"intelligence artificielle"

L'intelligence artificielle est une branche de l'informatique qui cherche à incorporer l'intelligence humaine dans des machines. On crée ainsi des systèmes permettant d'effectuer des tâches complexes pour lesquels on requière en temps normal cette intelligence humaine. Ils fonctionnent sur base d'algorithmes et de règles, avec un minimum d'intervention humaine [3, 2]. L'intelligence artificielle est donc une expression générique pour tout programme informatique ayant une forme d'intelligence humaine. Elle englobe à la fois le Machine Learning et le Deep Learning[2].

2.2 Le Machine Learning

Deepak Jakhar et Ishmeet Kaur expliquent que le Machine Learning regroupe « toutes les approches permettant aux machines d'apprendre à partir de données sans être explicitement programmées » [2]. Elles tournent sur base d'algorithmes et données et c'est via ces données et les informations traitées que les machines apprennent à prendre des décisions. Elles peuvent se modifier en étant exposées à plus de données. Par « learning » («apprentissage») on entend qu'elles font en sorte qu'il y ait le moins d'erreurs et que leurs prédictions soient le plus juste possible [2]. Le Machine Learning cherche à effectuer des tâches cognitives grâce à des modèles analytiques construits automatiquement, le but étant, par exemple, de détecter des objets, traduire du langage. Les ordinateurs peuvent trouver des idées cachées et des modèles complexes, tout cela, il faut le rappeler, sans être explicitement programmés. Toutes ces tâches sont effectuées via des algorithmes qui, de manière itérative, font un apprentissage à partir de données d'entraînement [3].

À partir d'un ensemble d'observation, un ensemble d'apprentissages, le Machine Learning a pour but d'obtenir une fonction de prédiction, la construction de cette fonction constitue l'apprentissage, ou l'entraînement du modèle [4]. Selon Ted Dunning, figure de proue du Machine Learning, un algorithme de Machine Learning doit avoir principalement les cinq qualités suivantes : il doit être facilement déployé, robuste, transparent (lorsque les performances d'une application dotée d'un Machine Learning se dégradent, l'algorithme doit le détecter le plus vite possible), adéquat aux compétences disponibles, être proportionnel

en temps et énergie investie par rapport aux bénéfices, enfin il doit être performant [4]. Par ailleurs, on juge la qualité d'un algorithme de Machine Learning par rapport à sa capacité à généraliser les associations qu'il a apprises durant la phase d'entrainement à de nouvelles observations [4].

Il existe trois types de Machine Learning : à apprentissage supervisé, non supervisé et par renforcement [3]. Il y a plusieurs familles d'algorithmes, comportant de nombreuses variantes , on a les modèles de régression, les algorithmes basés sur les instances, les arbres de décisions, les méthodes bayésiennes et les réseaux de neurones artificiels (ANN) [3]. Le choix de l'algorithme dépend d'abord du type de problème à résoudre [4].

L'apprentissage supervisé est la forme de Machine Learning la plus courante [4]. Dans ce type d'apprentissage, on dispose d'un ensemble de données d'apprentissage, avec en entrée des exemples (la variable x) et des réponses étiquetées (ou des valeurs de la variable cible y) en sortie. On entraine le modèle, c'est-à-dire qu'une phase d'apprentissage est effectuée, durant laquelle on observe les associations entre les données d'entrée et de sortie. On peut ainsi calibrer les paramètres du modèle. Lorsque l'entrainement du modèle est un succès, on peut l'utiliser avec de nouvelles données d'entrée (variable x) pour prédire la variable cible y [3, 4]. Cet apprentissage se subdivise encore en deux catégories. La régression, qui consiste à prédire une valeur numérique, la variable cible est une variable quantitative[4], et la classification consistant à attribuer le résultat d'une prédiction à une classe, une catégorie, la variable cible est donc ici qualitative [3, 4].

À l'inverse de l'apprentissage supervisé, dans l'apprentissage non supervisé, les données d'apprentissage ne sont pas étiquetées [3, 4]. Le but est que le système retrouve des informations structurelles par lui-même afin qu'il regroupe les exemples fournis en entrée par catégorie, c'est ce qu'on appelle le clustering (partitionnement). Le but peut être aussi de faire une réduction de dimension c'est-à-dire de projeter des données d'un espace avec une haute dimension vers un espace avec une faible dimension [3]. Ce système présuppose une notion de distance ou de similarité entre les observations [4].

Enfin, il y a l'apprentissage par renforcement pour lequel on ne donne pas de paires d'entrée et de sortie mais on fournit au modèle une description de l'état du système en cours, un objectif et une liste d'actions autorisées et de contraintes imposées par l'environnement. Une fois le modèle de Machine Learning lancé, on le laisse expérimenter par lui-même « en utilisant le principe d'essai et d'erreur pour maximiser une récompense » [3].

2.3 Le Deep Learning

On retrouve au sein du Machine Learning ce que l'on appelle les réseaux de neurones artificiels (artificial neural networks ou ANN), il s'agit d'une structure flexible pouvant être modifiée selon divers contextes et être ainsi utilisée dans les trois types de Machine Learning [3]. Le Deep Learning est justement une branche du Machine Learning comprenant des algorithmes imitant les réseaux de neurones d'un cerveau humain, on parle ainsi de réseau de neurones artificiels. À l'instar du cerveau, le réseau compare une nouvelle information à celles dont il dispose en stock afin de leur donner un sens, ainsi, il déchiffre, étiquette et assigne ces informations à la catégorie adaptée. Si on parle de « Deep » Learning, d'apprentissage « profond », c'est en raison du nombre de couches constituant le réseau. Le Deep Learning constitue donc une branche du Machine Learning. Il existe trois couches dans ce type de réseau. Il y a la couche d'entrée, recevant les données d'entrée, la couche de sortie, avec les résultats, et la couche cachée qui extrait les modèles depuis les données. Un réseau de neurones artificiels profond aura ainsi plus d'une couche cachée, une architecture profondément imbriquée, des neurones avancés, il peut donc effectuer des opérations avancées comme des convolutions [3]. Le Deep Learning fonctionne bien en particulier sur de grands volumes de données non structurées. Il est également plus précis que le Machine Learning. Il est cependant plus couteux à mettre en œuvre et nécessite par essence un énorme volume de données [2]. Mais des modèles de Machine Learning peu profonds peuvent être parfois supérieurs au Deep Learning, notamment dans le cas où les données d'entrainement sont peu disponibles ou lorsque les entrées de données sont de faibles dimensions [3]. Ainsi, grâce au Machine Learning et au Deep Learning, on peut notamment, entre autres applications, effectuer du traitement de langage naturel. Le traitement de langage naturel, ou Natural Language Processing (NLP), est un champ du Machine Learning qui a pour but d'analyser et extraire de précieuses informations depuis un texte [1].

3 Application de l'analyse sémantique dans la recherche historique

3.1 Outils et intérêt de l'analyse sémantique en histoire

3.2 Le problème de la sémantique des langues anciennes

3.3 Quelques exemples d'application sur des thématiques modernes

Références

[1] Arianna Di Bernardo, Simone Poetto, Pietro Sillano, Beatrice Villata, Weronika Sójka, Zofia Piętka-Danilewicz, and Piotr Pranke. Latin writing styles

analysis with machine learning : New approach to old questions. *CoRR*, abs/2109.00601, 2021.

[2] Deepack Jakhar and Ishmeet Kaur. Artificial intelligence, machine learning and deep learning : definitions and differences. *Clinical and experimental dermatology*, 45(1) :131–132, 2020.

[3] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3) :685–695, 2021.

[4] Pirmin Lemberger, Marc Batty, Médéric Morel, and Jean-Luc Raffaëlli. *Big Data et Machine Learning-3e éd. : Les concepts et les outils de la data science*. Dunod, 2019.

[5] Shmuel Liebeskind and Chaya Liebeskind. Deep learning for period classification of historical hebrew texts. *Journal of Data Mining & Digital Humanities*, 2020, 2020.

[6] Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray. Gasc : Genre-aware semantic change for ancient greek. *CoRR*, abs/1903.05587, 2019.