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 [6]. 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 [6]. 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 [8]. 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 [4, 3]. 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[3].

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 » [3]. 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 [3]. 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 [4].

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 [5]. 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 [5]. 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 [5].

Il existe trois types de Machine Learning : à apprentissage supervisé, non supervisé et par renforcement [4]. 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) [4]. Le choix de l'algorithme dépend d'abord du type de problème à résoudre [5].

L'apprentissage supervisé est la forme de Machine Learning la plus courante [5]. 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 [4, 5]. Cet apprentissage se subdivise encore en deux catégories. La régression, qui consiste à prédire une valeur numérique, la variable cible est une variable quantitative[5], et la classification consistant à attribuer le résultat d'une prédiction à une classe, une catégorie, la variable cible est donc ici qualitative [4, 5].

À l'inverse de l'apprentissage supervisé, dans l'apprentissage non supervisé, les données d'apprentissage ne sont pas étiquetées [4, 5]. 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 [4]. Ce système présuppose une notion de distance ou de similarité entre les observations [5].

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 » [4].

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 [4]. 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 [4]. 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 [3]. 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 [4]. 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

Aujourd'hui une énorme part de sources et de documents textuels se trouvent sous format numérique [2], derrière cette numérisation des textes sous format papier se trouve une volonté de conserver le patrimoine culturel et de faciliter son accès, non seulement à la communauté scientifique, mais aussi au grand public [6]. Cet état de fait donne un énorme corpus de travail pour des études requérant des outils de traitement de langage naturel.

Les outils de traitement du langage naturel peuvent améliorer qualitativement les recherches pour un historien, ils peuvent réduire la quantité d'information à consulter et rendre son travail plus efficace [2]. Par une analyse des données des

textes en langage naturel, c'est-à-dire du text mining, ou fouille de textes, on classifie et catégorise les textes, on recherche des informations (des informations documentaires non structurées du fait d'un besoin d'informations), on traite les changements dans les textes. Dans ce contexte, pour des besoins de classification et de recherche de l'information, on retrouve notamment la recherche d'informations sémantiques [2].

Les études de sémantique ont un certain intérêt en histoire, elles apportent un soutien à la recherche historique. On peut, par exemple, faire une recherche afin de voir les évolutions des sens d'un ou plusieurs mots à travers l'histoire ou les régions. Avec une telle recherche, on comble ainsi « le fossé lexical entre les langues modernes et anciennes » [6]. De plus, connaître les évolutions sémantiques en histoire permet de restreindre la recherche à certains sens en particulier [8].

Ainsi, la mise à disposition d'un nombre de documents historiques sous format numérique toujours plus croissant a poussé à mettre en application les méthodes et outils de traitement du langage naturel [6].

Le traitement du langage naturel est « une technique computationnelle qui permet l'analyse du langage » [1]. Il consiste en deux étapes, d'abord un prétraitement du texte. Ce prétraitement a pour but de nettoyer le texte, c'est-à-dire de traiter la ponctuation, supprimer les mots vides, corriger les erreurs d'orthographe, etc. La deuxième étape est le word embedding, ou plongement lexical, qui consiste à transformer le texte en objet mathématique sur lesquels une opération peut être effectuée. Il s'agit d'une étape nécessaire à l'analyse computationnelle. Pour ce faire, un vecteur est attribué à chaque mot ou phrase et ainsi les termes sont projetés dans un espace vectoriel [1]. À l'aide de ces vecteurs, on peut ensuite calculer la similarité entre les termes (avec la distance euclidienne ou encore la similarité cosinus) [1]. Les mots ayant une sémantique similaire ont des vecteurs similaires. il s'agit d'une approche basée sur les réseaux neuronaux. Les modèles les plus populaires de word embedding sont ceux issus de Word2Vec, avec notamment les algorithmes Skip-gram et continuous bag-of-word. Il s'agit dans les deux cas de réseaux neuronaux à deux couches, peu profonds et qui reconstruisent les contextes linguistiques des mots [6].

L'application d'outils de traitement du langage naturel aux textes historiques présente toutefois quelques difficultés. Ces textes comportent des propriétés linguistiques particulières, comme la grammaire, l'orthographe ou des abréviations qui ne sont pas standardisées. De plus, il y a accessoirement le problème de savoir la date à laquelle le texte a été écrit quand ce n'est pas précisé, ce qui peut être gênant lorsque l'on veut réaliser une étude sur l'évolution sémantique d'un ou plusieurs mots [6].

3.2 Le problème de la sémantique des langues anciennes

Un problème commun à toutes les langues quand on veut faire une étude diachronique sur un champ lexical c'est qu'une langue n'est jamais totalement « ancienne » ni « nouvelle » [6], il y a à tout moment une coexistence des sens originaux avec des nouveaux [8]. Il faut ajouter qu'il existe des tendances et des modes d'utilisation des langues, qui peuvent avoir un fort impact sur les champs lexicaux, y compris dans un espace de temps très restreint, parfois inférieur à cinquante ans [6].

Jusqu'à récemment, la recherche sur les changements sémantiques s'était principalement concentrée sur les langues modernes [7]. De ce fait, les outils de traitement du langage naturel ont pour la majorité été conçus pour les langues modernes, or les langues historiques sont différentes des langues modernes sur de nombreux aspects, ce qui rend problématique l'utilisation de tels outils pour ces langues [1].

Les langues anciennes présentent effectivement quelques difficultés non négligeables. Ainsi, le Grec ancien, par exemple, est une langue où les mots peuvent avoir un certain nombre de sens différents. Les changements sémantiques qu'il est possible de repérer peuvent dès lors être fort liés à ce problème de polysémie, ce qui accentue la difficulté de trouver notamment un moment où un sens nouveau apparaît [8].

Une autre difficulté avec les textes anciens est le fait que les dates de publications sont souvent inconnues, ce qui complique la tâche lorsque l'on travaille sur une évolution du sens des mots dans le temps, ainsi que lors du classement des textes historiques par période d'écriture, comme c'est le cas dans l'étude de Chaya Liebeskind et Shmuel Liebeskind, Deep Learning for Période Classification of Historical Texts. Il s'agit d'une étude ayant pour but de classifier un corpus [1] de texte en hébreu dans quatre périodes (XI^{e} siècle-fin XV^{e} siècle, XVI^{e} siècle, XVII^{e} siècle-XIX^{e} siècle, XX^{e} siècle à aujourd'hui). Sur base de cette classification, ils ont pu effectuer une recherche de changements sémantiques dans le temps. Dans leur cas, l'hébreu moderne soulève des problèmes, car il incorpore des mots de la Bible et de commentaires rabbiniques, l'usage de morphèmes d'hébreu biblique, de l'orthographe mishnique, de la prononciation séfarade, ainsi que d'expressions idiomatiques yiddish. Pour leur étude, ils ont fait usage de modèles de Deep Learning. Le Deep Learning s'avérait en effet être le plus efficace, car leur corpus comptant 1 406 208 mots, cela représentait un grand nombre de données, une situation dans laquelle excelle le Deep learning. Ils ont comparé trois modèles différents, à savoir les vecteurs de paragraphes (qui modélise l'espace thématique avec des vecteurs de paragraphes), le réseau convolutionel (convolutional neural network, ou CNN, un réseau neuronal à anticipation) et le réseau neuronal récurrent (recurrent neural network RNN, un

1. à savoir le corpus Responsa comptant 1 406 208 mots

réseau qui utilise « sa mémoire interne pour traiter des séquences arbitraires d'entrée »[6]). Leurs modèles ont utilisé des word-embedding de 300 dimensions qui ont été produits par l'algorithme skip-gram de Word2Vec, avec une fenêtre de cinq mots . Ils ont tout d'abord testé une classification via trois méthodes de Machine Learning conventionnelles (un Naive Bayes, un modèle linéaire et un Multi-Layer Perceptron). Ces trois méthodes conventionnelles se sont finalement montrées inférieures aux trois méthodes d'apprentissage profond qu'ils ont choisies. En effet, les méthodes conventionnelles avaient une efficacité qui dépendait beaucoup des étapes préalables, lors de la mise en place de l'ingénierie. Alors que l'avantage de l'apprentissage profond est qu'à partir de données d'entrée brutes, peu importe le domaine d'application, on fixe une importante quantité de caractéristiques à découvrir automatiquement. Ils ont donc par la suite opté pour des algorithmes d'apprentissage profond, appartenant à la catégorie des Machines Learning supervisés. De leurs trois modèles conventionnels, le modèle Naive Bayes a été le plus performant, mais de tous les modèles testés, les plus performants ont été les modèles Deep Learning CNN et RNN. Une fois la classification du corpus en quatre périodes réalisée, ils ont procédé à une analyse de changements sémantiques. Pour ce faire, ils ont adopté un protocole d'entrainement continu, dans lequel les embeddings de chaque période initialisent le modèle de la suivante. Ils ont entrainé les vecteurs de mots avec le package open-source Gensim. Pour analyser les changements sémantiques, ils ont d'abord comparé la similarité cosinus des mots des première et quatrième périodes, en retirant les mots apparaissant moins de 500 fois dans l'ensemble du corpus et ceux apparaissant moins de 50 fois dans chaque période. Ils ont par après comparé les mots voisins de ces mots cibles. Ils en déduisent que, lorsque les mots avaient des « comportements » différents, cela pouvait « provenir de changement de sens ou de changement de formulation dans le contexte » [6]. Ainsi, ils ont regardé l'évolution du sens des mots cibles par rapport à leurs mots voisins. Ils ont également calculé la similarité cosinus d'un même mot à travers les différentes périodes, en prenant la première comme période de référence, afin de retrouver des périodes de changement sans l'influence des mots voisins. Ils remarquent, alors, que la plupart des changements de sens ont eu lieu lors des deuxième (XVI^{e} siècle) et troisième périodes (XVII^{e}-XIX^{e}), ces changements s'expliquent par le fait que l'Hébreu redevient une langue couramment parlée à la fin du XIX^{e} [6].

On voit déjà ce que de tels outils nous permettent de réaliser en termes de recherches de changements sémantiques. De telles expériences ont également été mises en cours sur des corpus de textes grecs et latins. C'est ainsi ce qu'ont entrepris Valerio Perrone et al. dans deux études, une sur un corpus de textes grecs et une sur un corpus de textes latins en plus du grec. Mais dans ces études, Valerio Perrone et al. y intègrent un paramètre particulier.

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] Anna Glazkova, Valery Kruzhinov, and Zinaida Sokova. Automatic text processing for historical research. In *International Conference on Modern Information Technology and IT Education*, pages 135–146. Springer, 2017.

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

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

[5] 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.

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

[7] Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray. Lexical semantic change for ancient greek and latin. *Computational approaches to semantic change*, pages 287–310, 2021.

[8] 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.