Le nouveau Frankenstein: quand l'intelligence artificielle écrit le fiction.

Abstract

Ce travail tente d'explorer le nouvel univers de génération des oeuvres littéraires par les machines (ou plus précisément en collaboration avec les machines). Il aborde également le problème de l'attribution de l'auteur aux textes rédigés par des machines et comment cela pose un nouveau défi pour les chercheurs dans ce domaine. Je teste les similitudes et les différences entre les textes écrits par une machine et ceux écrits par un être humain et essaie de déterminer si les méthodes traditionnelles d'attribution d'auteur peuvent être utilisées pour détecter les textes écrites par une machine. Les résultats, les réflexions sur cette problématique ainsi que les propositions pour le futur travail sont discutés à la fin de cet article.

Introduction

“Words have more power than any one can guess; it is by words that the world’s great fight, now in these civilized times, is carried on.” Mary Shelley

Les mots ont le pouvoir de changer le monde. Ils peuvent inspirer les gens à faire des choses qu'ils n'auraient jamais envisagées auparavant. C'est à travers les mots que nous structurons nos pensées, nous apprenons notre histoire et notre culture. Nous utilisons les mots pour communiquer nos pensées et nos sentiments, pour exprimer nos besoins et nos désirs. Les mots sont tout simplement les briques de base de nos relations avec les autres, avec nous-mêmes et avec le monde qui nous entoure.

Les différents oeuvres littéraires nous accompagnent depuis notre plus jeune âge. La littérature est un vecteur important de la culture et permet aux individus de se connecter avec leurs racines. La connaissance culturelle apportée par la littérature revient à faire bouger ses propres conceptions identitaires, la littérature permet d'entrer et d'adopter de nouveaux points de vue, de se situer en reliance avec l'autre.

Les disputes concernant la propriété des textes existent depuis que les mots peuvent être possédés. L'attribution de l'auteur est l'un des problèmes les plus anciens et reste au centre de questionnement dans les recherches d'aujourd'hui. La problématique évolue avec le temps et l'arrivée des ordinateurs puissants apporte les nouvelles possibilités dans le traitement de cette ancienne problématique.

Également , un autre phénomène a apparue les dernières années et c'est arrivé d'un IA capable de générer du texte écrit d'une qualité telle qu'il est souvent difficile de le distinguer d'un texte écrit par un être humain. find citation 50 % rate humain distinguish gpt3 Dans les années à venir, il est indiscutable que cette technologie va progresser.

Dans le problème typique d'attribution d'auteur, le texte d'auteur inconnu est attribué à un auteur candidat, sur la base d'un ensemble d'auteurs candidats pour lesquels des échantillons de texte incontestés sont disponibles. Mais que se passe-t-il si l'auteur du travail n'est pas un être humain, mais une machine ? nous ne nous intéressons pas seulement à savoir qui est l'auteur du texte, mais également à déterminer si ce texte a été produit par une IA ou non.

Dès qu'on commence à explorer ce domain, les autres questions surgissent... Quelles sont les caractéristiques linguistiques qui peuvent être observées dans les textes écrits automatiquement ? Est-ce l'IA arrive à réproduire le style precis d'un auteur? Quelle aide apporten les méthodes d'attribution d'auteur?

Dans mon travail, je tente de trépondre à ces questions et de tester les similarités et les différences entre les textes écrits par l'IA ou par un écrivain.

- introduire correctement gpt3 dt la stylometrie

GPT3

Un modèle linguistique est un modèle mathématique qui calcule la probabilité qu'une séquence de mots se produise dans une langue donnée.

GPT 3 (acronyme de Generative Pre-trained Transformer 3) est un modèle de linguistique développé par la société OpenAI qui se fonde sur unlarge large corpus (Common Crawl) citation needed . Il s'agit d'un modèle linguistique comprenant 175 milliards de paramètres qui peut prédire les résultats à partir d'une simple requête en langage naturel sans avoir à mettre à jour ses poids. Tianyu Gao, Adam Fisch, Danqi Chen Making Pre-trained Language Models Better Few-shot Learners

"Les transformers" proposent un nouveau type d'architecture simple qui est composée de deux principales parties, un encodeur et un décodeur. L'encodeur prend le mot en entrée et le convertit en représentation vectorielle, qui est ensuite passée par un processus d'attention pour générer la prédiction du mot suivant.

Ce qui est "révolutionnaire" c'est le mécanisme d'attention qui permet à l'encodeur de transmettre au décodeur quel mot de séquence de l'entrée utiliser pour génerer un mot de sortie en parfaite correspondance.

Shazeer Noam, Parmar Niki Attention is all you need

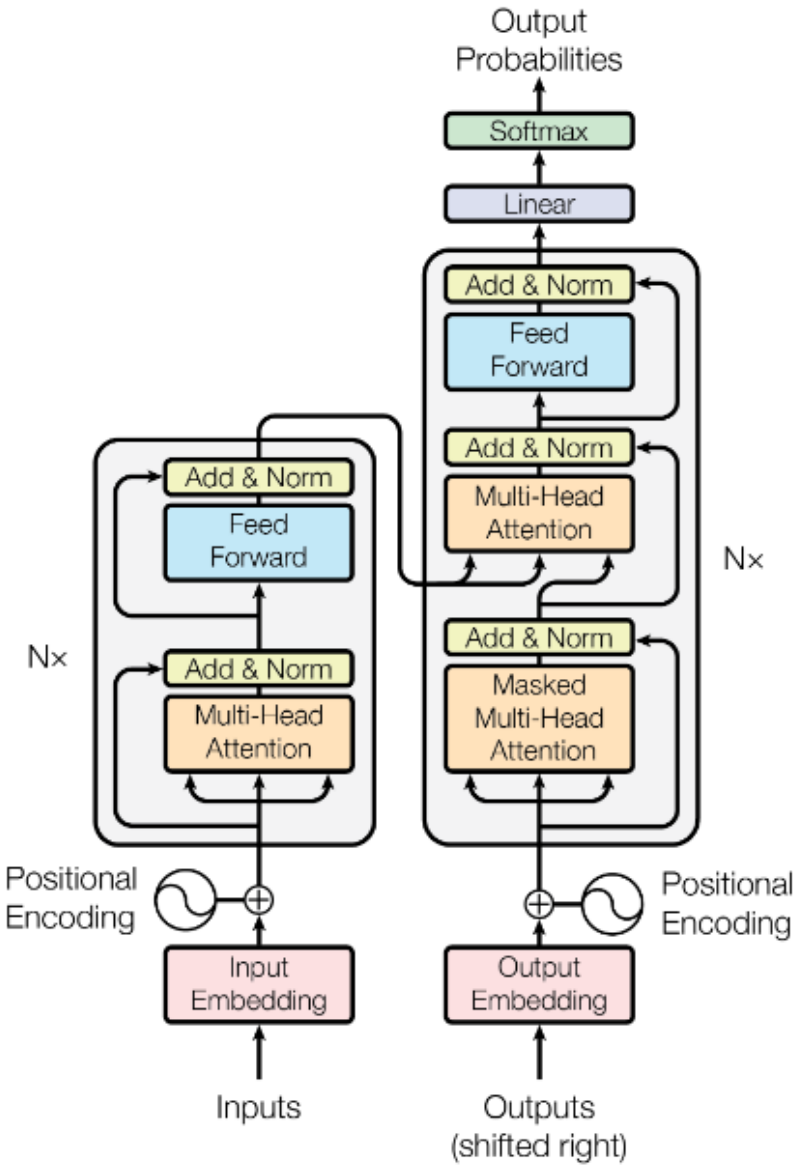

Figure 1: The Transformer - model architecture.

Language model prompting et fine-tuning.

GPT3 offre différents modèles/moteurs de langage, qui diffèrent en termes de fonctionnalités. Les modèles les plus importants sont : Davinci, Curie, Babbage et Ada. Davinci est considéré comme le modèle le plus performant et Ada est le plus rapide. Il es possible d'interagir avec le modèle par le biais de l'API et affiner les réglages du modèle via "fine-tuning" ou "prompting". Grâce à cela on peut obtenir de meilleurs résultats et ainsi améliorer l'efficacité du modèle.https://beta.openai.com/docs/guides/completion/introduction

- Fine-tuning: Le fine-tuning est un processus qui affine le modèle en sélectionnant une partie de donnes specifique au résultats souhaité. Cependant, un des inconvénients potentiels est la nécessité de créer un nouveau ensemble de données pour chaque nouvel tâche.

- Few-shots: On fournit au modèle quelques examples sans lui changer ses paramétres originaux. Le jeu de données dans ce cas contient entre 10 et 100 exemples.
- One-shot: Les mêmes principes s'appliquent qu'au «few-shots», à l'exception qu'un seul exemple est fourni au modèle.
- Zero-shot: Contraitrement aux deux modes précédents aucun exemple n'est donné juste une description en langage naturel.

Comme décrit largement dans l'article Language Models are Few-Shot Learners, GPT 3 donne les meilleures performances quand on lui fournit quelques exemples.

L'attribution d'auteur

La plupart des chercheurs pensent que chaque auteur a sa propre "empreinte linguistique" qu'on peut détecter dans l'écriture. ([Juola, 2008, p. 7] En réalité, le concept d'une empreinte linguistique est une métaphore qui n'est pas particulièrement utile et peut même être trompeuse. La valeur d'une empreinte digitale physique est que chaque échantillon est à la fois identique et exhaustif, ce qui signifie qu'il contient toutes les informations nécessaires pour identifier une personne. En revanche, tout échantillon linguistique, quelle que soit sa taille, ne fournit que des informations partielles sur son créateur.Coulthard Malcolm Author Identification, Idiolect, and Linguistic Uniqueness

Depuis le début du développement des méthodes de l'attribution d'auteur, de nombreuses techniques ont été proposées et finalement abandonnées en raison des controverses qu'elles suscitent. Dans certains domaines la question de l'exactitude est cruciale et sans une raison bien établie elle ne peut pas constituer une preuve admissible.
Rudman J. Rudman, “The state of authorship attribution studies a estimé qu'il y avait plus de mille caractéristiques qui sont utilisées dans les études d'attribution d'auteur, mais aucune n'est universelle et les meilleurs résultats sont obtenue avec utilisation de plusieurs méthodes. Juola Patrick Authorship Attribution

Les methods d'attribution d'auteur

Les propriétés clé du langage

Les relations entre différents éléments du texte ne sont pas aléatoires mais complexe ce pourquoi il est faire um modele informatique efficace. Juola Patrick Authorship Attribution

Caractéristiques linguistiques

1. Le vocabulaire en tant que caractéristique :

Le texte peut être vu comme une séquence de "tokens", qui sont les mots, les nombres et les signes de ponctuation. Cette méthode peut être utilisée sur à n'importe quel corpus sans aucune exigence supplémentaire, à l'exception de la disponibilité d'un tokenizer (c'est-à-dire un outil permettant de segmenter le texte en tokens).
Selon un accord general, ce qui caracterise au mieux le style d'un auteur se sont les mots fonctionnels, ou simplement les mots les plus frequents. Juola Patrick Authorship Attribution

De plus, dans cette catégorie on peut utiliser la richesse du vocabulaire comme un indice pour déterminer l'auteur d'un texte. Il reste intuitivement plausible que des personnes différentes aient des vocabulaires préférés différents (et des tailles de vocabulaire différentes). Cependant, cette méthode est facile à imiter et les données peuvent être facilement falsifiées. L'acuité de cette méthode est que la taille du vocabulaire dépend fortement de la longueur du texte.

2. Le syntaxe en tant que caractéristique :

Du point de vue sémantique les mots fonctionnels sont dépourvus de sens mais ils permet de décrire les relations entre les mots. Par exemple, le fonction principale du mot « de» est simplement d'établir une relation entre deux noms.

Une autre méthode qui s'appuie sur le syntaxe est POS ("Part of the speech"). En d'autres termes, les constructions syntaxiques préférées d'un auteur peuvent également être un indice de sa spécificité, de même que son utilisation de la ponctuation [25]. Les N-grammes (bigrammes, trigrammes, etc.) peuvent être utilisés pour capturer à la fois des informations lexicales et syntaxiques, ce qui permet une inférence plus précise. Cependant, les N-grammes peuvent également capturer des informations spécifiques au contenu, qui ne sont pas nécessairement indicatives du style. Efstathios Stamatatos A Survey of Modern Authorship Attribution Methods

3. Les autres caractéristiques

De multiple méthodes ont donc été proposé par les chercheurs. Certains analyse les documents par des séquences de caractères, plutôt que le vocabulaire. Cela peut aider à capturer les mots qui sont morphologiquement liés, qui seraient autrement manqués. Pour conclure, le tableau suivant apporte une vue d'ensemble des méthodes les plus utilisées.

En resumé:

TABLE 1. Types of stylometric features together with computational tools and resources required for their measurement (Brackets indicate optional tools.).

Features		Required tools and resources
Lexical	Token-based (word length, sentence length, etc.)	Tokenizer, [Sentence splitter]
	Vocabulary richness	Tokenizer
	Word frequencies	Tokenizer, [Stemmer, Lemmatizer]
	Word *n*-grams	Tokenizer
	Errors	Tokenizer, Orthographic spell checker
Character	Character types (letters, digits, etc.)	Character dictionary
	Character *n*-grams (fixed length)	–
	Character *n*-grams (variable length)	Feature selector
	Compression methods	Text compression tool
Syntactic	Part-of-speech (POS)	Tokenizer, Sentence splitter, POS tagger
	Chunks	Tokenizer, Sentence splitter, [POS tagger], Text chunker
	Sentence and phrase structure	Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser
	Rewrite rules frequencies	Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser
	Errors	Tokenizer, Sentence splitter, Syntactic spell checker
Semantic	Synonyms	Tokenizer, [POS tagger], Thesaurus
	Semantic dependencies	Tokenizer, Sentence splitter, POS tagger, Text Chunker, Partial parser, Semantic parser
	Functional	Tokenizer, Sentence splitter, POS tagger, Specialized dictionaries
Application-specific	Structural	HTML parser, Specialized parsers
	Content-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries
	Language-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries

Efstathios Stamatatos A Survey of Modern Authorship Attribution Methods A survey of modern authorship attribution methods, Stamatatos

Quelques mots autour l'oeuvre de Mary Shelley

Le style d'écriture de Mary Shelley est romantique et gothique. Il est caractérisé par une grande imagination, un sens de l'horreur et une atmosphère sombre et mystérieuse. Ses histoires mettent souvent en scène des personnages isolés et désespérés, luttant contre des forces surnaturelles ou contre leur propre nature.

Dans mon travail je vais utiliser uniquement ses deux oeuvres majeurs.

Frankenstein

Frankenstein est un roman d'épouvant écrit par Mary Shelley en 1818. Le roman met en scène le docteur Victor Frankenstein et sa créature, un monstre assemblé à partir de morceaux de cadavres. Frankenstein refuse de prendre responsabilité de sa créature et, en conséquence, le monstre devient une force destructive, tuant ceux que Frankenstein aime le plus. Le roman est considéré comme un classique de la littérature gothique et a inspiré de nombreuses œuvres de fiction et de films.

Le dernier homme

Le dernier homme est un roman de science-fiction apocalyptique et dystopique de Mary Shelley, publié pour la première fois en 1826. L'histoire se déroule en Europe à la fin du 21e siècle, ravagée par une mystérieuse maladie pandémique qui se propage rapidement à travers le monde entier, entraînant finalement l'extinction quasi totale de l'humanité.

Ces deux romans ont beaucoup en commun. Shelley remet en question la place de la personne dans la société. Frankenstein met avance les derives de la pensé scientique et de la créativités.
et Le dernier homme touche à l'extrême, à l'extanction de race humaine.
Phillips, Shannon Reanimating the Creature The Last Man as a Sequel to Frankenstein (1999). Masters Theses. 1512.

Methodologie de travail

Corpus d'évaluation

Je me suis concentré dans mon travail sur la fiction de Mary Shelley et ses deux œuvres majeures, "Frankenstein" et "The Last Man". Je voulais comprendre comment l'IA génère du texte et quelles similitudes nous pouvons trouver entre les textes automatiques et les œuvres littéraires. Pour ajouter une autre aspect de comparaison, j'ai ajouté un travail écrit par un fan de Mary Shelley que j'ai trouvé sur le site fanfiction.com.fanfiction link

Le site fanfiction.com est l'un des plus importants archives et forums de fanfiction au monde, où des auteurs et des lecteurs de fanfiction du monde entier se rassemblent pour partager leur passion. Mon critère de sélection pour le texte de fanfiction était la longueur du texte (20 000 mots et plus) et que l'histoire ait quelques points favoris. Je pense qu'ajouter ce autre corpus permet de voir si le style GPT 3 est complètement différent du style écrit par un humain ou pas.

Ce qui touche le pré-traitement des corpus, je n'ai pas éliminé les mots fonctionnels ni la ponctuation. J'ai simplement supprimé la page d'index et toutes les informations relatives au Projet Gutenberg, et ensuite j'ai tokenisé la totalité de texte.

La productions de textes par GPT3

Comme mentionné dans la partie théorique, GPT 3 est l'IA qui a été entraînée sur un grand nombre de textes (Common Crawl). Il s'agit d'une archive web qui consiste de pétaoctets de données collectées depuis 2011. Les œuvres qui se trouvent dans le domaine public font partie de cette archive, ce qui signifie que GPT3 a également été entraîné sur les œuvres de Mary Shelley et les connaît "par cœur".

Cet aspect, que j'ai completement négligé au début de mon projet m'a posé beaucoup de problèmes par la suite car je ne pouvais générer que des copies de l'œuvre de l'auteur. Après de nombreuses heures de recherches de solutions au problématique de "finetuning" je me suis tourné vers la solution de "prompting" par une invite manuelle. https://www.gwern.net/GPT-3 Selon la littérature scientifique, le prompting manuel donne les résultats les moins cohérents que les autres méthodes mais on capte bien la "créativité" du modèle.

Un autre problème que j'ai rencontré lors de la création de mon GPT3 corpus était la longueur du texte. Il est difficile de produire un long texte cohérent par le prompting manuel. OpenAI recommende d'utiliser la longueur maximale de 2048 tokens générés. Face à ce problème, j'ai décidé de "contourner" cette difficulté et de produire plusieurs histoires d'une longueur maximale de 2048 jetons et de les assembler dans le seul document. Ce n'est certainement pas la manière idéale de produire un texte, mais après avoir lu ces courts histoire j'ai pu découvrir qu'il sont fortement influencées par l'univers de Mary Shelley.

Les réglages pour obtenir le corpus GPT3:

- la temperature était réglé entre 0.7 - 0.9 (plus le chiffre est proche de 1, mois modèle est rigide)
- max.token : 2048
- engine = Davinci (modèle le plus performante de GPT3)
- les prompts :
  query = "Write long story in Mary Shelley style"
  query = "Write long award winning story in Mary Shelley's style"
  query = "Write long story about apocalypse in Mary Shelley's style"
  query = "Write long story about end of mankind in Mary Shelley's style"
  query = "Write long story about new science in Mary Shelley's style"
  query = "Write long story about man in Mary Shelley's style"
  query = "Write long story about woman in Mary Shelley's style"
  query = "Write long story about monsters in Mary Shelley's style"
  query = "Write long science fiction in Mary Shelley's style"

Les méthodes d'attribution d'auteur

Comme discuté dans la section théorique, il existe un vaste univers de méthodes stylométriques qui peuvent distinguer les textes de différents auteurs. Après plusieurs recherches, j'ai fait confiance au livre Real Python et à programming historian pour explorer les méthodes les plus classiques. Ce ne sont certainement pas les méthodes les plus sophistiquées et précises, mais proposent une première exploration des ces méthodes.
Principales méthodes utilisées: fréquence des mots fonctionnels, POS(partie du discours), test de Mendenhall de la longueur des mots et méthode du chi-carré de Kilgariff. Les descriptions des méthodes explorées sont rassemblées dans le Notebook Jupyter correspondant.

Resultats de mes recherches

Le test des stopwords montre que le mot «the» est le mot le plus fréquent dans les trois corpus. Les autres stopwords varient selon le corpus. Les mots fonctionnels les plus fréquents dans le corpus de Shelley sont «the», «of», «and». Les stopwords les

plus utilisés par GPT 3 sont «the», «and», «was» et enfin dans le Fanfiction «the», «he», «his». Les courbes montrent que Shelley et GPT3 ont beaucoup de similitudes.

Le test Mendellhall montre que les textes courbes générés par GPT3 sont beaucoup plus proches de Shelley que de Fanfiction. Il semble que GPT 3 génère plus de très courtes phrases et utilise moins de mots longs.

Il n'est pas surprenant que les résultats du test de vocabulaire montrent de nombreuses similitudes entre GPT 3 et le corpus de Shelley. GPT 3 génère des textes en utilisant le vocabulaire de l'auteur, ce qui explique pourquoi le test du chi 2 est beaucoup plus proche de Shelley que de Fanfiction. Le vocabulaire n'est qu'une partie du style de l'auteur, mais en ce qui concerne ce dernier, GPT3 semble être le plus performante.

Dans un test de catégorie grammaticale, on peut voir les similarités entre la fanfiction et GPT3. Les deux corpus ont les trois caractéristiques les plus fréquentes : NN : Nom, singulier, VBD : Verbe, temps passé, IN : Préposition ou conjonction subordonnée

A l'inverse : L'œuvre de Mary Shelly est composée de NN : Nom, singulier, IN : Préposition ou conjonction subordonnée DT : Déterminant

Le coefficient de Jaccard est une statistique utilisée pour évaluer la similarité et la diversité d'un ensemble d'échantillons.

Conclusion

Ce projet a été une plongée dans un grand océan d'inconnu. Il est né de mon amour pour la littérature et de ma grande curiosité pour découvrir comment, en utilisant l'IA, nous pouvons collaborer avec les machines et trouver les nouveaux moyens d'expression artistique. J'ai réalisé au cours de ce projet où sont mes intérêts mais aussi où sont mes limitations techniques.

Je n'ai pas pu me tenir complètement au plan que je m'étais fixé au début de ce projet en raison des difficultés rencontrées en cours de route. J'ai «perdu» des heures plongé dans la lecture des articles scientifiques en essayant de mieux comprendre le fonctionnement de GPT3 et passé encore plus de temps à y réfléchir. Ce travail est une exploration, pas une destination, et plus je passais de temps à réfléchir à la manière d'aborder ce problème, plus d'autres questions se posaient. Je suis conscient que mon approche manque de rigueur et, aujourd'hui, je vois beaucoup de choses que j'aurais pu faire différemment si j'avais recommencé à zéro.

Pour finir, je citerai Marvin Minsky, car plus je réalise de projets informatiques, plus ce terrain de jeu me procure de la joie... et un peu de folie.

“To constrain the behavior of a program precisely to a range may be very hard, just as a writer will need some skill to express just a certain degree of ambiguity. A computer

is like a violin. You can imagine a novice trying first a phonograph and then a violin. The latter, he says, sounds terrible. That is the argument we have heard from our humanists and most of our computer scientists. Computer programs are good, they say, for particular purposes, but they aren't flexible. Neither is a violin, or a typewriter, until you learn how to use it."

Marvin Minsky, "Why Programming Is a Good Medium for Expressing Poorly-Understood and Sloppily-Formulated Ideas" 1967