



AUTOMNE 2020

À LA RECHERCHE D'UN MODÈLE EN LANGUE ANCIENNE

L'IA AU SECOURS DE L'ENSEIGNEMENT DU LATIN

DAMIEN CAVALERI

MSPRO35

Haute Ecole Pédagogique Vaudoise



PROBLÉMATIQUE

L'intérêt fondamental de ce projet est de chercher à démontrer le potentiel de l'intelligence artificielle en tant que support de l'enseignement d'une langue ancienne telle que le latin. Car, si, pour les langues modernes, le volume de matière de difficulté variable nécessaire à l'acquisition de celles-ci est facilement disponible (grâce, notamment, au fait qu'il existe des locuteurs natifs), cela est loin d'être le cas pour les langues anciennes. Dans l'optique de remédier à ce manque, nous cherchons à mettre en place un modèle de réseau neuronal de type transformateur, fondé sur l'architecture du célèbre GPT-2, et à l'entraîner sur un corpus de textes latins en vue de générer des phrases latines non seulement grammaticalement et syntaxiquement correctes, mais aussi cohérentes. Le projet s'oriente essentiellement sur un double axe de recherche, à savoir : relever les apports potentiels de l'intelligence artificielle en termes d'acquisition d'une langue et proposer des perspectives sur l'application d'un générateur textuel profond en tant qu'outil didactique, d'accompagnement et de validation des acquis.

INTRODUCTION

Quae nova nunc tellus medio dimota profundo

Que représente pour vous la Sicile ? Quelles images et quels sentiments vous inspire-t-elle ? Une île, la mer, l'activité tectonique, la poésie ? Si oui, vous vous entendriez bien avec le générateur intelligent ayant généré cet hexamètre dactylique. C'est, en effet, une de ses réponses à notre demande de description de l'île méditerranéenne et il y en a d'autres, plus intéressantes encore. Ici, il semble avoir choisi de parler de la création de celle-ci ou, plutôt, de son élévation depuis les abysses de la mer, tout en privilégiant une forme versifiée dont les structures métrique, syntaxique, syntagmatique et phonétique feraient pâlir de jalousie n'importe quel latiniste aguerri. Mais passons sur le fait que ce vers représente potentiellement le premier vers latin cohérent jamais créé par une intelligence artificielle et intéressons-nous plutôt à l'objet de ce présent travail.

Les langues anciennes présentent une difficulté particulière tant pour ceux qui l'apprennent que pour ceux qui l'enseignent en cela qu'elles ne possèdent plus de locuteurs « natifs » à partir desquels un usage authentique peut être inféré. Cela implique qu'il est, aujourd'hui, impossible d'appréhender une langue telle que le latin au travers du questionnement et de l'imitation d'un modèle vivant. Nous proposons, dans ce projet, d'explorer une voie de remédiation à ce manque en nous appuyant sur l'intelligence artificielle. Concrètement, il s'agit, à partir d'un corpus de textes latins régularisés, d'entraîner un réseau neuronal à générer, tel un locuteur natif, une langue de qualité. En effet, si une intelligence artificielle parvient à déterminer des relations probabilistes claires à l'interne et à l'externe de phrases latines, elle devrait être capable de formuler des phrases cohérentes et, surtout, *inédites* en fonction de ce qu'on lui demande. En d'autres termes, le but que nous poursuivons est de recréer le modèle qui fait défaut.

Un tel modèle représenterait une avancée certaine pour l'apprentissage et pour l'enseignement d'une langue ancienne comme le latin. Mis à disposition des enseignants, il

permettrait de les soutenir dans leur effort de 'latiniser' leur enseignement, d'employer la langue plus librement en réduisant le nombre d'inconnues -liées à l'inexistence de modèle et, donc, au fait que l'on ne sait tout simplement pas comment exprimer certaines choses- et ainsi de réduire la distance perçue entre eux-mêmes et la langue. Mis à disposition des apprenants, il permettrait de les accompagner dans leur construction et appropriation des différents savoirs, de les amener à travailler sur des textes inédits, voire à analyser ses erreurs, le cas échéant, et à réfléchir sur les processus qui l'y auraient mené. Ainsi tant les qualités que les défauts du modèle revêtent un intérêt didactique, puisque les uns comme les autres nous poussent à la réflexion sur nos propres connaissances et sur l'usage de la langue latine.

Nous proposons, dans ce dossier, de présenter succinctement la mise au point et la phase de préparation du générateur dont il est question, puis de montrer comment nous nous en sommes servis en tant qu'outil pédagogique, tout au long de l'entraînement du modèle, en nous servant des imperfections de ce dernier comme leviers d'apprentissage. Cela nous permettra, en outre, d'illustrer la manière dont nous avons impliqué les élèves d'une classe de maturité dans le processus de conception de l'outil. Au terme de ce rapport, nous aurons, ainsi, parcouru quelques exemples d'usage à visée pédagogique du générateur entraîné et ouvert la porte vers de nouvelles voies d'application intéressantes pour l'avenir.

LE GÉNÉRATEUR

Dans cette section, nous nous bornons à décrire le modèle utilisé et le processus de traitement des données d'apprentissage (*l'exemple à suivre*). Notons, toutefois, que la mise en place du générateur textuel latin présenté ici a nécessité un nombre élevé de tentatives avec des méthodes d'encodage, des architectures et des modèles différents ainsi que plusieurs centaines d'heures de traitement du corpus d'entraînement, ce que nous avons choisi de laisser de côté dans ce dossier, mais que la brièveté de ce dernier ne rend pas.

Modèle du réseau neuronal

L'architecture du modèle utilisé pour notre expérience est reprise de celle du modèle GPT-2-345M (*Generative Pre-trained Transformer 2*), développé par OpenAI en 2019 et mis à disposition sur Github¹. Notons que cette architecture est, elle-même, déjà une adaptation d'un modèle de transformation. Elle présente, néanmoins, l'avantage d'être capable de générer (en anglais) des textes attribuables à un être humain grâce, d'une part, à la taille du modèle (plus de 1,5 milliard de paramètres pour le plus gros) et, d'autre part, à la quantité de données d'apprentissage (40GB). La particularité de GPT-2, donc, est d'être un énorme transformateur. Toutefois, craignant une utilisation malhonnête de son modèle, l'équipe créatrice n'a pas rendu public le code permettant d'entraîner le réseau neuronal, elle a, en revanche, publié celui du modèle et un rapport technique décrivant le procédé d'entraînement².

Dans notre cas et sans compter la phase préparatoire, la mise sur pied du générateur actuellement utilisé dans ce travail et l'obtention de résultats probants a nécessité, au moment de boucler ce présent dossier, la mobilisation d'environ 12GB de mémoire vive sur GPU pendant un peu plus 4000 heures. Cela correspond à 672 itérations sur un ensemble d'environ 13'000'000 mots ou parties de mot pour un total de 65MB de texte. Notons aussi

¹ Cf. <https://github.com/openai/gpt-2>

² Cf. Alec RADFORD, Jeffrey WU, Rewon CHILD, David LUAN, Dario AMODEI, Ilya SUTSKEVER (OpenAI, 2019).

que, bien que nous ayons repris l'architecture de GPT-2, nous n'avons cependant pas fait usage du modèle préentraîné.

Quelle est la particularité de GPT-2 ?

Le modèle est une adaptation de transformateur et est entraîné de manière semi-supervisée à générer un texte de longueur fixée au préalable en fonction d'un élément textuel de départ (une amorce) ou à partir de rien.

Un transformateur est un modèle permettant la génération d'une séquence en fonction d'une séquence initiale, comme, par exemple, lors d'une traduction. Ce type de réseau est généralement composé de deux parties fondamentales : un composant d'encodage de la séquence initiale et un composant de décodage permettant de générer une nouvelle séquence. Le principe de l'encodage est de transformer une séquence reçue (mots, parties de mot, caractères ou nombres) en un vecteur (tableau de coordonnées) stockant les propriétés de la séquence. Le composant de décodage permet de faire l'inverse et donc de générer une nouvelle séquence vectorielle en fonction de la première. Dans les faits, un transformateur est composé de plusieurs blocs de transformations successifs ou non linéaires. Ainsi, un composant d'encodage ou de décodage peut comporter plusieurs encodeurs ou décodeurs. Une autre particularité d'un modèle de transformation est qu'il tient compte du contexte interne de la séquence au moyen d'un mécanisme d'attention (tous les mots, parties de mot ou caractères de la séquence sont analysés en fonction de tous les autres de la même séquence)³. Ce mécanisme d'attention peut être bidirectionnel, comme c'est le cas pour BERT⁴, ou masqué, comme pour GPT-2. Cela correspond à la partie de la séquence prise en compte lors de la génération d'une nouvelle séquence. Ainsi, un système bidirectionnel tient compte de l'ensemble d'une phrase, alors qu'un système masqué ne tient compte que d'une partie.

L'architecture de GPT-2-345M est, comme nous l'avons dit, une adaptation du modèle de transformation, mais présente tout de même des différences notables. En effet, le but d'un transformateur est, avant tout, de *transformer* une séquence en une autre (traduction), alors que celui de GPT-2 est de générer la suite d'une séquence en fonction du début de celle-ci. Cela signifie que GPT-2 doit créer une nouvelle séquence en fonction d'une amorce et non adapter une séquence entière en une autre. Dans cette optique, l'équipe d'OpenAI a choisi de sortir complètement le composant d'encodage des blocs individuels de transformation et de le placer au début du cycle. Cela implique que le GPT-2 comporte une succession de blocs de décodeurs initiée par un bloc d'encodage double (*embedding*). Chaque bloc de transformation (décodeur) est constitué de deux couches de normalisation, un perceptron multicouche (2

³ Ce qui n'est, par exemple, pas le cas avec un réseau neuronal récurrent doté de cellules LSTM ou GRU, puisque chaque cellule traite une séquence dans l'ordre. Dans ce cas, il est seulement possible d'inverser l'ordre (cellules bidirectionnelles) pour analyser la même séquence dans un autre ordre.

⁴ Cf. Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE, Kristina TOUTANOVA (Google, 2019).

couches convolutionnelles) et une couche d'attention masquée (masque la fin de la séquence, la partie devant être successivement prédite). Le modèle utilisé dans ce travail possède 24 blocs de transformation et se clôt sur une dernière couche de normalisation suivie d'une couche convolutionnelle. Les valeurs sortantes sont ensuite normalisées exponentiellement. Cela permet d'obtenir une répartition des probabilités sur 1 de l'élément suivant directement la séquence entrante. Afin d'assurer un apprentissage progressif et, surtout, adaptatif, l'erreur du modèle est minimisée au moyen d'*Adam (Adaptive moment estimation)*, qui est une variante de la descente de gradient stochastique tenant compte du « moment » (la moyenne mobile exponentielle) de la fonction objective en une position donnée. La mesure d'erreur employée est celle de l'entropie croisée des prédictions et des phrases d'apprentissage. Elle permet d'évaluer la capacité du modèle à générer la suite d'une phrase par rapport à la phrase réelle. Cependant, l'utilisation de cette mesure, en dehors du domaine mentionné ci-dessus, présente un certain risque d'évaluation incorrecte de la qualité des générations. En effet, dans les langues humaines, l'amorce d'une phrase ne prédétermine, en réalité, pas nécessairement la suite de celle-ci. En effet, une phrase commençant, par exemple, par « Lorsque j'étais en Sicile, ... » ne possède pas qu'une seule manière d'être poursuivie, bien que nous sachions qu'elle introduit la description d'un phénomène concomitant à la présence du sujet en Sicile. Ainsi, cette phrase d'exemple pourrait se terminer tant par « je me suis cassé une jambe » que « j'ai fait de la plongée » sans pour autant être incorrecte. Et pourtant, si nous calculions l'entropie croisée entre ces deux phrases, nous verrions qu'elle est relativement élevée. Cet exemple illustre le rôle et le défaut inhérent de la mesure de perte employée, puisque celle-ci indique le degré de maîtrise du corpus d'entraînement sans pour autant récompenser la diversité des générations⁵. Pourtant la qualité concrète du modèle, du moins celle que nous recherchons, est, précisément, de s'émanciper du corpus d'entraînement. Cela pose la question de la pertinence de l'utilisation de l'entropie croisée pour optimiser la perte du modèle.

Fonctionnement de GPT-2

Nous avons mentionné, plus haut, que l'apprentissage était « semi-supervisé », cela signifie qu'il n'est pas nécessaire de fournir manuellement les *bonnes réponses* au modèle lors de la phase d'entraînement. En effet, le modèle est entraîné à prédire ce qui suit une séquence en fonction des séquences que nous lui avons fournies. Les *labels* (bonnes réponses) sont, par conséquent, contenus dans les séquences d'entraînement. Le principe de fonctionnement de GPT-2 est donc de prédire le prochain mot ou caractère ou partie de mot en fonction d'une séquence fournie. Puis, lors de la génération d'un texte plus long, de tenir compte de sa dernière prédiction (ajoutée à la séquence initiale) pour générer un nouvel élément de phrase. Ce processus se répète ensuite itérativement dans la séquence jusqu'à la génération d'un

⁵ Ping CAI, Xingyuan CHEN, Peng JIN, Hongjun WANG, Tianrui LI (2020) proposent d'utiliser un classificateur pour mesurer la perte et ainsi récompenser la diversité, avec des résultats intéressants, mais nous ne l'avons pas implémentée dans le modèle dont il est question dans ce projet.

token de fin de ligne. Cette méthode permet au modèle de créer des textes complets et, surtout, cohérents, puisqu'il tient compte de l'entier d'une séquence pour en générer la suite. Le modèle entraîné dans ce travail comprend 345 millions de poids (345M). Les poids ou paramètres sont des valeurs numériques attribuées aux relations (*edges*) entre neurones à l'interne du modèle qui représentent les critères pondérés sur lesquels se fonde celui-ci pour générer une prédiction. Pour note, le plus grand modèle GPT-2 comprend 1,5 milliard de poids (GPT-2-1.5B) et le plus grand GPT-3 en comporte 175 milliards. Mais les faire tourner nécessite une quantité de mémoire et une puissance de calcul que nous n'avons pas.

Dans notre cas de figure et étant donné que nous ne pouvions pas profiter de la formule préentraînée, il a été nécessaire d'adapter le modèle pour un réentraînement depuis zéro sur la langue latine.

En résumé, le réseau neuronal apprend à générer en latin un élément de séquence suivant une séquence donnée en imitant les séquences latines contenues dans le corpus d'apprentissage.

Préparation et nettoyage des données d'apprentissage

Entraîner un transformateur nécessite un très grand nombre de données de très bonne qualité. Pour ce faire, nous avons récupéré, relu et corrigé autant que possible l'entier des textes latins disponibles sur <https://www.thelatinlibrary.com/> et sur <http://www.perseus.tufts.edu/hopper/> répondant aux conditions suivantes :

- L'œuvre est postérieure au 8^{ème} siècle avant J.-C. et antérieure au 6^{ème} siècle ap. J.-C.
- L'œuvre possède des phrases suivies et complètes (cela exclut les recueils de fragments).

Si une bonne part du nettoyage textuel des sources peut être automatisée, il reste, malheureusement, une part irréductible de corrections manuelles, plus ou moins importante selon la qualité des textes. Pour l'essentiel, ces défauts prennent les formes suivantes :

- Variantes orthographiques
 - Propres au latin classique (*optimus/optumus*)
 - Propres aux manuscrits sources (*tertius/tercius, Italus/Ytalus, obt-/opt-*)
- Incohérences d'apparat critique (à l'intérieur du texte)
- Erreurs de numérisation (automatisée)
 - Identification de lettre ('b' au lieu de 'h', 'rn' au lieu de 'm', etc.)
 - Caractères spéciaux

Notons que ces défauts s'observent rarement dans les générations. Il est, néanmoins, probable qu'ils influent négativement sur la qualité des prédictions, puisqu'ils empêchent l'identification de certains mots et faussent les fréquences de ces derniers.

Certaines variantes orthographiques très courantes, telles que *-re*, *-vere*, *-runt* pour *-verunt* (3^{ème} personne du pluriel parfait actif), *-re* pour *-ris* (2^{ème} personne du singulier passif) ou encore *capiundum* pour *capiendum* n'ont pas été modifiées.

Une fois que la qualité du corpus a été jugée satisfaisante, nous avons mis en place un dictionnaire des éléments existant dans les séquences contenues dans le corpus et leur avons attribué une valeur numérique proportionnelle à leur fréquence respective dans le même corpus (*tokenization*). Dans notre cas, nous avons fait usage de la technique de compression par encodage des paires de bytes (*Byte Pair Encoding, BPE*). Le nombre d'entrées dans ce dictionnaire a été limité à 50'256 de manière à réduire l'usage de mémoire et à accélérer l'encodage⁶. Le principe de l'encodage par paires de bytes (BPE) est de relever les combinaisons de bytes (caractères) les plus fréquentes et d'allouer un byte (l'entrée dans le dictionnaire) à chacune de ces combinaisons. Lors de nos essais successifs, nous avons noté que cette méthode d'encodage des séquences offrait les meilleurs résultats en matière de progression de l'apprentissage.

Compte tenu des limitations de temps, de matériel adéquat et de financement suffisant, notre expérience porte sur un corpus de 65MB de textes latins. Ce qui est déjà énorme, bien que très peu en comparaison des 40GB de GPT-2. De manière générale, il semble exister une corrélation directe entre la quantité de données d'apprentissage, la taille du transformateur et l'augmentation de la qualité des phrases générées.

⁶ Cf. Elizabeth SALESKY, Andrew RUNGE, Alex CODA, Jan NIEHUES, Graham NEUBIG (2020).

L'APPLICATION

Nous voici maintenant au cœur du projet, à savoir l'application du générateur textuel dans un cadre pédagogique. Dans cette section, nous survolons, d'abord, les différents intérêts et bénéfiques que peut apporter l'utilisation d'un tel outil, puis, dans un second temps, nous présentons quelques voies d'application réelle du modèle, dont, notamment, un exercice proposé à des élèves de 1^{ère} année de Maturité OS Latin.

L'intelligence artificielle est appelée à jouer un rôle de plus en plus important dans un avenir proche. Aujourd'hui déjà, sans que nous en soyons forcément conscients, elle touche à toutes les facettes de nos existences. C'est pourquoi, même si elle devait n'avoir aucune utilité dans l'enseignement d'une langue ancienne, il serait primordial de l'introduire d'une manière ou d'une autre et le plus tôt possible dans le cursus scolaire obligatoire. Car prendre du retard dans ce domaine équivaut à condamner les générations futures au silence et à les livrer au pouvoir de ceux-là seuls qui la maîtriseront sans partage⁷.

Les avantages et bénéfiques de l'usage de l'intelligence artificielle sont innombrables et visibles, du moins, nous l'espérons, au travers de ce présent dossier ainsi que de notre mémoire. En effet, quel meilleur outil pourrait-on souhaiter pour mieux saisir les rouages internes de la langue latine ? Quel meilleur moyen pour permettre à chacun, enseignant, élève ou apprenant, de discerner la qualité de la langue, de nettoyer les manuels et précis de grammaire des pseudorègles et de lever les incertitudes ? L'intelligence artificielle nous permet, aujourd'hui, d'atteindre un niveau de compréhension de la langue latine inégalable au cours du dernier millénaire. Il serait une erreur de s'en priver.

Aujourd'hui enfin, après plus d'un millénaire de transmission orale et textuelle d'une langue employée par des gens, certes fort érudits et plus érudits, probablement, que nous-mêmes nous ne le serons jamais, mais qui ne vécurent pas à l'époque où le latin était une langue vernaculaire, il est possible, pour la première fois, d'étudier la langue latine directement

⁷ Cf. Francesc PEDRO, Miguel SUBOSA, Axel RIVAS, Paula VALVERDE (UNESCO, 2019).

auprès des auteurs antiques. Il est, en effet, surprenant que le contenu des précis de grammaire change au fil des siècles alors que le latin classique lui-même n'est pas censé s'être modifié depuis près de 2'000 ans. Il semble, de plus, que ces modifications s'accroissent à partir du 19^{ème} siècle.

Tous ces changements, qu'ils soient bons ou mauvais, sont les conséquences de décisions humaines, des décisions qui ont soit apporté une correction à une erreur antérieure soit introduit une nouvelle erreur. L'usage de l'intelligence artificielle, pour autant qu'elle ne soit pas elle-même biaisée par le développeur, permet d'éviter au latiniste de tomber dans les écueils de la faillibilité humaine et de jouir de règles, d'usages et de formes non biaisées. Elle permet, en effet, d'étudier la composition, l'ordre, la facture et le style d'une phrase, d'émettre des règles fondées sur un usage purement antique (ou de la période de son choix), de discerner les pondérations des différents composants langagiers et de représenter un usage non tributaire des biais inhérents à un enseignement humain.

La syntaxe et la grammaire latines diffèrent dans une mesure relativement importante de celles des langues modernes. Néanmoins, il peut être parfois difficile pour un enseignant de langues anciennes privilégiant une approche vivante de différencier correctement les expressions latines des expressions vernaculaires latinisées (gallicisme, anglicisme, etc.). Dans ce contexte, un modèle intelligent de discrimination textuelle (juste/faux) donnera à chacun les outils pour évaluer et améliorer son expression tant orale qu'écrite. De la même manière, un modèle de génération de qualité permettra d'obtenir des expressions correctes selon le type de texte étudié. Ce dernier modèle offre également un autre avantage, plus terre-à-terre : la possibilité de générer des textes de facture classique conditionnés par une amorce. Cette amorce peut revêtir diverses formes et viser différents objectifs. Notons, tout de même, que, dans notre cas précis, nos tentatives de pilotage des générations ont montré que cibler précisément un thème au moyen de l'amorçage pouvait se révéler relativement ardu.

Dans l'optique de démontrer l'intérêt de l'intelligence artificielle en tant qu'outil pédagogique ainsi que ses capacités et ses limites, nous présentons ici deux cas d'usage du générateur entraîné. Le premier exemple illustre la génération d'un texte continu, dont le style a été piloté par une amorce adéquate. Bien que les générations longues comportent encore souvent des défauts notables en matière de cohérence, cela permet d'illustrer la capacité du modèle à être piloté selon des critères prédéfinis. Le dernier exemple offre un exercice présenté à des élèves du secondaire II dans le cadre d'une révision grammaticale en OS Latin.

Pour note, les générations représentées ci-dessous sont le produit d'une sélection des générations les plus adaptées. Pour obtenir des phrases adéquates, il faut souvent tester plusieurs amorces différentes et adapter ces dernières en fonction des prédictions obtenues.

Composition latine : générer un texte de style et de contenu ciblés

Grâce à l'inférence conditionnelle, il est possible de générer un texte en fonction d'une amorce de longueur variable. Il en découle la possibilité de compléter conditionnellement une

œuvre fragmentaire en tenant compte de l'ensemble de celle-ci et pour autant qu'elle soit de taille conséquente. Dans la même veine, -et cela représente une piste d'application que nous envisageons de mettre en pratique dans notre classe ces prochains mois- il est possible de se servir du générateur comme outil de comparaison dans le cadre d'un exercice de composition latine. Ici, l'intérêt concret de l'usage du modèle réside dans le fait d'amener les élèves à analyser, évaluer et à justifier les propositions d'ajout de celui-ci et à chercher à les situer dans une certaine logique. Cette méthode correspond, à peu de choses près, à l'application de l'outil illustrée dans le second exemple ci-dessous. Elle diffère, néanmoins, de cette dernière dans la mesure où elle place une emphase plus sur le contenu et moins sur la langue elle-même -sans pour autant l'exclure. Au travers de cet exemple, nous tentons d'illustrer un cas de génération d'exercice ciblé sur un auteur et sur un contenu précis.

L'exemple proposé reprend une partie d'un paragraphe du *Satyricon* de Pétrone situé au début de l'œuvre. Il s'agit, ici, pour les élèves comme pour le générateur, de proposer, en quelques phrases, une suite à la fin abrupte du paragraphe.

Pour résumer le contenu de ce dernier en quelques mots, Encolpe (narrateur) et Ascylte essaient de revendre un manteau qu'ils ont volé. Pour ce faire, ils se rendent, en fin de journée, au marché. Deux personnes, un « bouseux » et une « femmelette », montrent de l'intérêt pour le manteau. Encolpe trouve un air familier au premier, mais ne parvient pas à se rappeler où il l'aurait déjà vu. Pendant que celui-ci essaie le manteau, Ascylte lève les yeux et voit quelque chose qui le laisse sans voix. La suite manque. L'objectif de l'exercice est de finir ce paragraphe. Notons, tout de même, que le générateur possède l'avantage de déjà connaître l'ensemble de l'œuvre.

Amorces : variations de l'amorce suivante

ueniebamus in forum deficiente iam die, in quo notauius frequentiam rerum uenialium, non quidem pretiosarum sed tamen quarum fidem male ambulans obscuritas temporis facillime tegeret. cum ergo et ipsi raptum latrocinio pallium detulissemus, uti occasione opportunissima coepimus atque in quodam angulo laciniam extremam concutere, si quem forte emptorem splendor uestis posset adducere. nec diu moratus rusticus quidam familiaris oculis meis cum muliercula comite propius accessit ac diligentius considerare pallium coepit. inuicem ascyltos iniecit contemplationem super umeros rustici emptoris, ac subito exanimatus conticuit.

Générations : sélection de séquences de longueur limitée suivant l'amorce fournie et obtenues après 672 époques.

Une dispute au marché, indice de variété : 4

cinaedus ille maesto statim fletu commotus et : 'o quam maribus manum infunditis. dominam tuam, pertinax, tota nocte, basiabo.' 'ut qui ab hac eiectus sis, inquit rusticus, repente custodiaris ?'

La « femmelette » (le générateur applique, ici, le terme à un homme, alors que dans le récit de Pétrone, c'est une femme) se met soudainement à pleurer et se lance dans une plainte à propos des mains baladeuses que subissent les hommes, puis se tourne vers le « bouseux » et lui dit, pour se venger, qu'elle embrassera toute la nuit la maîtresse de celui-ci. L'autre lui répond « Pour que tout d'un coup elle prenne ta défense alors qu'elle t'a chassé ? ».

Un complice trahi, indice de variété : 12

procedentibus deinde longius marito et muliere coactis indignatione gitona tormentis offerentibus, ecce ancilla advenit et : 'quid ? ego, inquit, non sum digna ut collega, sed uictoris consilii sim, eiusdem etsi rei oblita ?'

Alors que le mari et la femme (le « bouseux » et la « femmelette ») s'éloignent, indignés, et veulent mettre Giton (autre personnage, compagnon d'Encolpe et d'Ascylte) à la torture, voici qu'une servante arrive et clame : « Quoi ? Je ne suis pas digne, moi, d'être votre associée (à la combine du manteau ?), mais juste bonne à avoir les bonnes idées, même si j'ai oublié de quoi il s'agissait ?

Ces deux exemples amusants permettent d'illustrer la capacité du générateur d'attribuer à un extrait donné un registre stylistique et thématique, ici, pétronien, en fonction de l'œuvre source. En conséquence, bien que le sens global des générations souffre parfois de certaines incohérences, le modèle se sert d'une grammaire, d'une syntaxe et d'un lexique cohérents et fait intervenir des thèmes et des personnages propres aux sources identifiées. Notons, toutefois, qu'il arrive très fréquemment, surtout si l'indice de variété est bas, que le générateur produise un texte très similaire, voire, parfois, identique à l'original (celui de Pétrone). Dans ce cas, il peut être utile de remplacer certains mots de l'amorce par leurs synonymes.

Les deux exemples ci-dessus ont été sélectionnés parmi plusieurs centaines de générations et, pour être honnête, sont d'une qualité supérieure à la moyenne.

Révision grammaticale : générer un exercice de révision grammaticale

Ce second exemple illustre un exercice proposé à des élèves pendant le semestre d'automne 2020. Cet exercice s'est inscrit dans un programme de révision et d'harmonisation grammaticale et avait pour objectif de permettre aux élèves de consolider leurs connaissances en matière de grammaire latine. L'intérêt, ici, était de faire usage des faiblesses du générateur et d'amener les élèves non seulement à identifier, corriger et justifier les erreurs du texte, mais aussi à réfléchir sur le potentiel (autant positif que négatif) de l'usage d'un générateur textuel intelligent et de son impact possible dans la société actuelle. L'exercice dont il est question se trouve en annexe (exercice du 16.11.2020).

Contexte de mise en œuvre de la tâche

- Niveau : OS Latin, Secondaire II, 1^{ère} de Maturité, 4^{ème} année de latin, 5 élèves.
- Objectif du plan d'études concerné : « En 1^{ère} année, harmonisation, révision, voire apprentissage des connaissances acquises en morphologie et en syntaxe ; compléments en fonction des besoins suscités par les lectures. »
- Cadre : révision grammaticale

Présentation de la tâche

La tâche possède deux objectifs essentiels d'apprentissage : premièrement, de réviser l'ensemble des thèmes grammaticaux revus depuis le début de l'année et, secondement, d'analyser et d'évaluer les stratégies linguistiques employées par un tiers. Elle comporte deux parties : la première contient neuf phrases produites par le générateur au cours de son entraînement et la seconde une invitation à évaluer le travail de celui-ci. Les erreurs observées dans les phrases générées sont grammaticales, syntaxiques, lexicales et de l'ordre de la cohérence. Ces phrases ont été sélectionnées, au préalable, par l'enseignant en fonction de leur adéquation avec les concepts traités lors des cours précédents et dans le cadre de la révision grammaticale. Chaque phrase comporte toujours soit une soit aucune erreur, mais peut faire intervenir plusieurs thèmes de grammaire.

Lors de l'introduction de la tâche aux élèves, nous avons mis l'emphase sur la nécessité pour eux de prendre une posture scientifique ou critique et de chercher, avant tout, à comprendre le cheminement *logique* (emphase sur la nature logique du raisonnement) derrière les erreurs du modèle. Notons que les erreurs présentes dans les phrases de l'exercice sont des erreurs rencontrées relativement fréquemment en classe. Cet exercice nous permet ainsi d'amener les élèves, tout en prenant une posture d'autorité active (domaine affectif), à repérer ces erreurs potentiellement familières (domaine cognitif), à verbaliser les stratégies qu'emploie un tiers pour y parvenir (domaine métacognitif) et à identifier les méthodes de remédiation (domaines cognitif et métacognitif). L'objectif était, par conséquent, d'amener les élèves à réfléchir sur leurs propres pratiques en matière d'analyse de phrase et de les pousser progressivement vers une prise de conscience des processus de pensée qu'eux-mêmes mettent parfois en œuvre devant une phrase et vers l'exercice d'un contrôle régulateur sur ceux-ci.

Il a, en outre, été extrêmement intéressant de remarquer que les élèves ont eu beaucoup plus de facilité à identifier les stratégies menant à une erreur et à y remédier qu'à repérer l'erreur elle-même.

Quant à l'évaluation finale, en seconde partie, elle poursuit le double objectif, d'une part, de confirmer le statut du lecteur – à savoir, de celui d'évaluateur critique et actif – et, d'une autre, de réviser le complément du comparatif (à l'ablatif) et l'usage des adverbes. Pour note, les élèves ont jugé, ici, que le travail du générateur était *satis bene* ('satisfaisant').

Consignes

- Lisez les phrases latines générées par le générateur
- Identifiez les erreurs probables, puis corrigez et justifiez-les du point de vue du modèle
- Évaluez la qualité globale des générations analysées

Thèmes principaux intervenant dans chaque phrase

1. Comparatif et déclinaison des adjectifs neutres de la seconde classe
2. Cohérence du propos (le Rhône est une rivière, pas un lac)
3. Infinitives
4. Prépositions suivies de l'ablatif et déclinaison des adjectifs de la première classe
5. Participes présents
6. Conjugaison
7. Infinitives
8. Prépositions suivies de l'ablatif et *cum* suivi du subjonctif
9. Concordance des temps

Évaluation : Comparatif et adverbes

Bilan

Les élèves ont montré un intérêt particulier pour l'intelligence artificielle ainsi que pour son application à la langue latine et ont fait preuve d'un investissement notable dans l'exécution de la tâche. Celle-ci possédait, en outre, l'avantage de faire appel à des habiletés tant cognitives et affectives que métacognitives⁸ tout en les amenant, au travers de la prise d'une posture d'autorité évaluatrice par rapport à un tiers (le générateur), à assumer un rôle pleinement actif dans la construction de leur savoir. En effet, l'usage du générateur de même que le *modus operandi* présenté ci-dessus ont favorisé un climat d'interactivité dialogique⁹, qu'il est souvent difficile d'atteindre dans un cadre de révision grammaticale, et ont contribué à introduire une conscientisation des pratiques, raisonnements ou processus de pensée mis en œuvre face à un problème de déchiffrement d'une phrase latine, d'abord au travers d'un processus de verbalisation, puis d'évaluation et, enfin, de régulation.

La conjonction de ces éléments fait que l'exercice a été plutôt apprécié par les élèves et que nous prévoyons de nous en servir plus fréquemment, sous forme réduite, comme rituel de début de classe.

⁸ Selon la taxonomie de VERMUNT et VERLOOP (1999).

⁹ Nous faisons référence, ici, à la « communication dialogique interactive » telle que la définissent SCOTT et AMETLLER (2007).

PERSPECTIVES

Dans cette dernière partie, nous survolons rapidement les limites et les voies d'amélioration possibles pour l'avenir en vue d'améliorer la qualité des générations et offrons quelques réflexions finales sur l'application d'un tel outil dans un contexte pédagogique.

Globalement et compte tenu de la nouveauté de la technologie, du temps et du matériel à disposition, les résultats obtenus par le modèle de génération textuelle sont relativement bons. Les limites notables que nous avons rencontrées et qui sont illustrées dans les deux exemples présentés à la section précédente sont, essentiellement, de l'ordre de la cohérence et du pilotage des générations.

L'existence d'incohérences, en particulier pour les longues générations, est un défaut qui se retrouve jusque dans le modèle GPT-2-1.5B¹⁰, même s'il est moins notable, et dont il semble, par conséquent, difficile de se défaire. OpenAI a fait le pari d'augmenter la taille du modèle et ce pari a payé, puisque les incohérences se sont réduites. Mais jusqu'où faudra-t-il encore augmenter le nombre de paramètres pour parvenir à un texte indifférentiable d'un texte humain ? Il se peut également que l'architecture de transformateur ne soit pas la réponse à la question épineuse du traitement du langage humain.

Toutefois, il est probable que le risque d'incohérences soit exacerbé par les défauts qualitatifs et quantitatifs du corpus d'entraînement. En effet, bien que nous ayons cherché autant que possible à nettoyer et à régulariser celui dont nous nous sommes servis lors de l'apprentissage, il est certain que des erreurs subsistent et influent négativement sur les résultats finaux. En ce sens, la régularisation, l'uniformisation et l'optimisation du nettoyage de corpus représentent une voie d'amélioration sûre, quoiqu'ingrate, et font partie de nos priorités pour l'avenir. Il faut aussi noter que la taille réduite du corpus (65MB au lieu des 40GB d'OpenAI)

¹⁰ Nous pouvons supposer que ce défaut est soit imperceptible soit inexistant avec le modèle GPT-3-175B. Malheureusement, il est difficile de le vérifier étant donné que le code n'est pas en libre accès. Cf. <https://openai.com/blog/openai-api/>.

augmente massivement le risque de surentraînement (*overfitting*) du modèle¹¹. Pour pallier le défaut de quantité, il serait nécessaire d'élargir le période de sélection des textes (entre le 8^{ème} siècle avant J.-C. et le 6^{ème} siècle après J.-C.) afin d'englober les écrits médiévaux, humanistes, modernes et contemporains. Mais cela ouvre la voie à tout un panel de nouvelles variantes orthographiques.

Un autre défaut majeur, que nous avons mentionné au début de ce travail et qui influe potentiellement sur la cohérence des générations, est le risque élevé de faux négatifs dû à l'emploi de l'entropie croisée pour minimiser la perte du modèle. En effet, celle-ci ne récompense pas la diversité des générations. L'erreur (perte) d'une génération est évaluée en fonction de la distance entre celle-ci et la phrase initiale. Cela a pour effet de rendre l'évaluation des générations plus rigide qu'il n'est souhaitable et d'augmenter fortement les risques de surapprentissage (*overfitting*).

Quant aux difficultés liées au pilotage des générations, il faut noter que nous disposons, actuellement, de trois méthodes pour ce faire : l'intuition, le tâtonnement et l'augmentation du nombre de critères contenus dans l'amorce. Il est, en effet, difficile de définir précisément quels sont les critères d'attribution à tel ou tel registre de telle ou telle partie de phrase ou combinaison de parties. Dans ce cas, le réseau neuronal est une sorte de boîte noire dont la logique peut parfois nous échapper. Pour pallier ce problème, certains chercheurs proposent d'adjoindre un modèle de classification au modèle de génération afin de prétraiter les données entrantes et, ainsi, d'en extraire les concepts clés sous une forme adaptée au générateur. Il s'agit, dans ce cas, de coupler deux types de modèles lors de l'entraînement et de la génération¹². Une autre solution offrant des résultats probants serait d'annoter (labelliser) les données d'entraînement en fonction de ce que nous désirons obtenir (par exemple : 'grammaire', 'déclinaisons', '1^{ère} conjugaison', etc.)¹³. Cela représente une piste intéressante, en particulier pour une application pédagogique. Elle est, toutefois, le fruit d'un travail d'édition manuelle du corpus d'entraînement de plusieurs années. Notons, cependant, que, de façon générale et de la même manière que pour un être humain, une amorce contenant un nombre plus élevé de critères produira une génération de meilleure qualité qu'une amorce plus courte.

Nous avons, ainsi, relevé cinq voies essentielles d'amélioration pour l'avenir :

- Cohérence
 - Augmentation du nombre de paramètres
 - Optimisation et uniformisation du traitement de corpus d'entraînement
 - Augmentation de la taille du corpus

¹¹ GPT-2 a été entraîné moins d'une époque sur 40GB de données. Cela signifie que nous devrions, théoriquement et au maximum, entraîner le nôtre 615x plus longtemps sur 65MB pour obtenir des résultats similaires, d'où un risque plus élevé de surentraînement.

¹² Cf. Virapat KIEUVONGNGAM, Bowen TAN, Yiming NIU (2020).

¹³ Cf. Varun KUMAR, Ashutosh CHOUDHARY, Eunah CHO (2020) sur l'élémentarisation de la littérature scientifique.

- Récompenser la diversité des générations du modèle au moyen d'une mesure de l'erreur réalisée par un discriminateur¹⁴ plutôt qu'à partir de l'entropie croisée
- Pilotage
 - Annotation supplémentaire du corpus par classificateur¹⁵ ou manuelle¹⁶

Les exemples présentés ont permis d'introduire l'usage d'un générateur textuel dans un cadre scolaire. Évidemment, le domaine d'utilisation est bien plus large et n'est, en réalité, limité que par notre imagination. Néanmoins, nous pensons qu'il est important de souligner que, même si les phrases générées ne sont pas encore parfaites, elles permettent tout de même d'envisager des perspectives réjouissantes pour l'avenir. En effet, il n'est pas difficile de voir que l'intelligence artificielle peut amener énormément à la promotion, à l'apprentissage et à l'enseignement du latin. Car, malgré les 2'000 années qui nous séparent, il nous est enfin donné la possibilité d'étudier le latin directement auprès des locuteurs natifs classiques. Ce qui était encore impensable il y a 10 ans est devenu une réalité et cette réalité nous la vivons aujourd'hui. Car c'est bien dans une nouvelle ère qu'entre l'enseignement du latin : celle du dialogue avec les anciens, celle de la renaissance d'une langue. *Exit* les interprétations douteuses, *exit* les vernacularismes, *exit* les à-peu-près, *exit* les règles grammaticales de convenance ! De la même manière que pour les langues modernes, les enseignants ainsi que les apprenants pourront faire appel à un « locuteur natif » pour comprendre, l'observer, l'analyser et l'imiter afin de parvenir à ce que nous visons tous : la maîtrise de la langue latine.

¹⁴ Cf. Ping CAI, Xingyuan CHEN, Peng JIN, Hongjun WANG, Tianrui LI (2020).

¹⁵ Cf. Virapat KIEUVONGNGAM, Bowen TAN, Yiming NIU (2020).

¹⁶ Cf. Varun KUMAR, Ashutosh CHOUDHARY, Eunah CHO (2020).

BIBLIOGRAPHIE

Bases de données et ressources en ligne

<https://arxiv.org/>, Cornell University, Ithaca, NY, USA.

<http://www.perseus.tufts.edu/hopper/>, TUFTS University, Medford, MA, USA.

William L. CAREY *et al.*, <https://www.thelatinlibrary.com/>, USA.

<https://fr.wikisource.org/>, Wikimedia Foundation, USA.

<https://github.com/>, Github Inc., USA.

<https://openai.com/>, OpenAI, CA, USA.

Général

Francesc PEDRO, Miguel SUBOSA, Axel RIVAS, Paula VALVERDE (2019) *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*, Working Papers On Education Policy, UNESCO.

Transformation

Alec RADFORD, Jeffrey WU, Rewon CHILD, David LUAN, Dario AMODEI, Ilya SUTSKEVER (2019), *Language Models are Unsupervised Multitask Learners*, rapport technique (GPT-2), OpenAI, San Francisco CA, USA.

Dzmitry BAHDANAU, Kyunghyun CHO, Yoshua BENGIO (2015), *Neural Machine translation by Jointly Learning to Align and Translate*, ICLR, San Diego CA, USA.

Elizabeth SALESKY, Andrew RUNGE, Alex CODA, Jan NIEHUES, Graham NEUBIG (2020), *Optimizing Segmentation Granularity for Neural Machine Translation*, *Machine Translation*, 34, pp. 41-59.

Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE, Kristina TOUTANOVA (2019), *BERT : Pre-training of Deep Bidirectionnal Transformers for Language Understanding*, Google AI Language, Mountain View CA, USA.

Tom B. BROWN, Benjamin MANN, Nick RYDER, Melanie SUBBIAH, Jared KAPLAN, Prafulla DHARIWAL, Arvind NEELAKANTAN, Pranav SHYAM, Girish SASTRY, Amanda ASKELL, Sandhini AGARWAL, Ariel HERBERT-VOSS, Gretchen KRUEGER, Tom HENIGHAN, Rewon CHILD, Aditya RAMESH, Daniel M. ZIEGLER, Jeffrey WU, Clemens WINTER, Christopher HESSE, Mark CHEN, Eric SIGLER, Mateusz LITWIN, Scott GRAY, Benjamin CHESS, Jack CLARK, Christopher BERNER, Sam McCANDLISH, Alec RADFORD, Ilya SUTSKEVER, Dario AMODEI (2020), *Language Models are Few-Shot Learners*, rapport technique (GPT-3), OpenAI, San-Francisco CA, USA.

Usage et implémentation de GPT-2

Alex BOYD, Raul PURI, Mohammed SHOEYBI, Mostofa PATWARY, Bryan CATANZARO (2020), *Large Scale Multi-Actor Generative Dialog Modeling*, NVIDIA & University of California, Irvine CA, USA.

Aran KOMATSUZAKI (2019), *One Epoch Is All You Need*, <https://arxiv.org/abs/1906.06669>.

Chaitra HEGDE, Shrikumar PATIL (2020), *Unsupervised Paraphrase Generation Using Pre-Trained Language Models*, pré-impression, Fidelity Investments, Boston MA, USA.

Ehsan HOSSEINI-ASI, Bryan McCANN, Chien-Sheng WU, Semih YAVUZ, Richard Socher (2020), *A Simple Language Model for Task-Oriented Dialogue*, Salesforce Research, San Francisco CA, USA.

Jinyi HUN, Maosong SUN (2020), *Generating Major Types of Chinese Classical Poetry in a Uniformed Framework*, Tsinghua University, Beijing, CN.

Nils KOEBIS, Luca MOSSINK (2020), *Creative Artificial Intelligence – Algorithms vs. humans in an incentivized writing competition*, University of Amsterdam, Amsterdam, NL.

Pedro HENRIQUE MARTINS, Zita MARINHO, André F. T. MARTINS (2020), *Sparse Text Generation*, Instituto de Telecomunicações & Priberam Labs & Institute of Systems and Robotics & Unbabel, PT, USA.

Ping CAI, Xingyuan CHEN, Peng JIN, Hongjun WANG, Tianrui LI (2020), *Distributional Discrepancy : A Metric for Unconditional Text Generation*, Southwest Jiaotong University & Leshan Normal University, CN.

Varun KUMAR, Ashutosh CHOUDHARY, Eunah CHO (2020), *Data Augmentation using Pre-trained Transformer Models*, Alexa AI (Amazon), Seattle WA, USA.

Virapat KIEUVONGNGAM, Bowen TAN, Yiming NIU (2020), *Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2*, Rockefeller University, New York NY, USA.

Langue latine

Damien CAVALERI (2019), *Tellis : L'épopée de Guillaume Tell suivie d'une étude sur la facture de l'hexamètre dactylique épique*, Université de Lausanne, Lausanne VD, CH.

Klaus THRAEDE (1978), *Der Hexameter in Rom : Verstheorie und Statistik*, C.H. Beck, München, DE.

Léo DE NEUBOURG (1986), *La base métrique de la localisation des mots dans l'hexamètre dactylique*, Paleis der Academien, Bruxelles, BE.

Lucio CECCARELLI (2008), *Contributi per la storia dell'esametro latino*, vol. I & II, Herder, Roma, IT.

Pédagogie

Jan D. VERMUNT, Nico VERLOOP (1999), Congruence and friction between learning and teaching, *Learning and Instruction*, 9(3), pp. 257–280.

Lev VYGOTSKI (1934), *Pensée et Langage*, traduction de Françoise SEVE (1997), La Dispute, Paris.

Phil SCOTT, Jaume AMETLLER (2007), Teaching Science in a Meaningful Way : Striking a Balance Between ‘Opening up’ and ‘Closing down’ Classroom Talk, *The School Science Review*, 88 (324), pp. 77-83.

ANNEXES

Exercice du 16.11.2020.....p. 25

EXERCITIA

ARTIFICIALIS INTELLIGENTIAE SENTENTIAE (post iterationes XXX - XL)

I. [...] an tibi putas satis esse nihil, nisi quod mortale maius sit ?

.....

II. Romani, cum lacus Rhodani pontes rescidissent, captivos viderunt.

.....

III. Dicebat etiam ipse epicureus eum non recte fecisse, si illum damnasset.

.....

IV. [...] venire se nolle dixit, ne ad se ab impudicum aliquid adlatum esse diceretur.

.....

V. Si et nos de nostra vita senserimus, spectabimus potius ad mortem illudentiorum.

.....

VI. Hocine existimas, tibi fore famae, si tu ad scelera confugerint ?

.....

VII. Lacedaemoniorum exemplo aiebat se rei publicae conscium fore.

.....

VIII. Hannibal autem Syracusis potitus, cum in portum Atheniensium obsideretur et ipsam classem adductam in suam potestatem redegisset, [...]

.....

IX. Ipsa mulier ait mihi, 'mihi decem nummos dari volo, ut hunc mulierem esse dicerem.'

.....

Valetne artificialis intelligentia ?

Minime Paulo minus opinione Satis bene Paulo melius opinione Optime