Qu’est-ce que le NLP ?

Le traitement automatique du langage naturel (NLP) est un sous-domaine de l’informatique et de l’intelligence artificielle (IA) qui utilise le machine learning pour permettre aux ordinateurs de comprendre et de communiquer en langage humain.

Le NLP permet aux ordinateurs et aux appareils numériques de reconnaître, comprendre et générer du texte et de la parole en combinant la linguistique computationnelle (la modélisation du langage humain basée sur des règles) avec la modélisation statistique, le machine learning et l'apprentissage profond.

La recherche en NLP a joué un rôle clé dans l'essor de l'IA générative, que ce soit pour les capacités de communication des grands modèles de langage (LLM) ou pour la capacité des modèles de génération d'images à comprendre des requêtes. Le NLP est déjà intégré dans la vie quotidienne de nombreuses personnes, en alimentant des moteurs de recherche, en activant des chatbots pour le service client via des commandes vocales, ou encore des systèmes GPS à reconnaissance vocale et des assistants numériques comme Alexa d'Amazon, Siri d'Apple et Cortana de Microsoft.

Le NLP joue également un rôle croissant dans les solutions d’entreprise destinées à rationaliser et automatiser les opérations métier, augmenter la productivité des employés et simplifier les processus métier.

Avantages du NLP

Il facilite la communication et la collaboration entre l'homme et la machine en permettant l'usage quotidien du langage naturel. Cela offre des avantages dans de nombreux secteurs et applications :

Automatisation des tâches répétitives

Amélioration de l'analyse des données et des informations

Amélioration de la recherche

Génération de contenu

Automatisation des tâches répétitives

Le NLP est particulièrement utile pour l'automatisation totale ou partielle de tâches telles que le support client, la saisie de données et le traitement de documents. Par exemple, les chatbots alimentés par le NLP peuvent traiter les requêtes courantes des clients, libérant ainsi les agents humains pour des cas plus complexes. Dans le traitement des documents, les outils de NLP peuvent automatiquement classer, extraire les informations clés et résumer le contenu, réduisant ainsi le temps et les erreurs associés au traitement manuel des données. Le NLP facilite également la traduction linguistique, en convertissant un texte d'une langue à

une autre tout en préservant le sens, le contexte et les nuances.

Analyse des données améliorée

Le NLP améliore l'analyse des données en permettant l'extraction d'informations à partir de données textuelles non structurées, telles que les commentaires des clients, les publications sur les réseaux sociaux et les articles de presse. Grâce aux techniques de fouille de texte, le NLP peut identifier des modèles, des tendances et des sentiments qui ne sont pas immédiatement apparents dans de vastes jeux de données. L'analyse des sentiments permet l'extraction de qualités subjectives telles que des attitudes, des émotions, le sarcasme, la confusion, la suspicion, etc. d'un texte. Cette fonction est souvent utilisée pour acheminer les communications vers le système ou la personne la plus apte à fournir la réponse adéquate.

Ainsi, les entreprises peuvent mieux comprendre les préférences des clients, les conditions du marché et l'opinion publique. Les outils de NLP peuvent également catégoriser et résumer de grandes quantités de texte, permettant ainsi aux analystes de repérer plus facilement les informations clés et de prendre des décisions fondées sur des données de manière plus efficace.

Recherche améliorée

Le NLP améliore la recherche en permettant aux systèmes de comprendre l'intention qui se cache derrière les requêtes des utilisateurs, ce qui permet d'obtenir des résultats plus précis et plus pertinents dans leur contexte. Plutôt que de s'appuyer uniquement sur la correspondance de mots-clés, les moteurs de recherche alimentés par le NLP analysent le sens des mots et des phrases, ce qui facilite la recherche d'informations même lorsque les requêtes sont vagues ou complexes. Cela améliore l'expérience utilisateur, que ce soit pour les recherches sur le web, l'extraction de documents ou les systèmes de données d'entreprise.

Puissante génération de contenu

Le NLP utilise des modèles de langage avancés pour générer des textes qui ont l'air d'avoir été écrits par la main de l'homme dans divers contextes. Des modèles préentraînés, tels que GPT-4, peuvent générer des articles, des rapports, des contenus marketing, des descriptions de produits et même des textes créatifs à partir des instructions fournies par les utilisateurs. Les outils alimentés par le NLP peuvent également contribuer à l'automatisation de tâches comme la rédaction d'e-mails, la création de publications sur les réseaux sociaux ou la rédaction de documents juridiques. En tenant compte du contexte, du ton et du style, le NLP s'assure que le contenu généré est cohérent, pertinent et en accord avec le message souhaité, tout en économisant du temps et des efforts dans la création de contenu sans compromettre la qualité.

Approches du NLP

Le NLP combine la puissance de la linguistique informatique avec les algorithmes de machine learning et l'apprentissage profond. La linguistique informatique utilise la science des données pour analyser le langage et la parole. Elle inclut deux principaux types d'analyse : l'analyse syntaxique et l'analyse sémantique. L'analyse syntaxique détermine le sens d'un mot, d'une expression ou d'une phrase en analysant la syntaxe des mots et en appliquant des règles de grammaire préprogrammées. L'analyse sémantique utilise les résultats syntaxiques pour tirer le sens des mots et interpréter leur signification dans la structure de la phrase.

L'analyse des mots peut prendre deux formes. L'analyse de dépendance examine les relations entre les mots, par exemple en identifiant les noms et les verbes, tandis que l'analyse par constituants construit un arbre d'analyse (ou arbre syntaxique) : une représentation enracinée et ordonnée de la structure syntaxique d'une phrase ou d'une chaîne de mots. Ces arbres d'analyse sous-tendent des fonctions des traducteurs automatiques et des systèmes de reconnaissance vocale. Idéalement, cette analyse rend le résultat, que ce soit un texte ou un discours, compréhensible à la fois pour les modèles NLP et pour les utilisateurs humains.

L'apprentissage auto-supervisé (SSL) est particulièrement utile pour soutenir le NLP, car ce dernier nécessite de grandes quantités de données étiquetées pour former les modèles d'IA. Comme l'annotation de ces jeux de données est un processus fastidieux, requérant l'étiquetage manuel effectué par des humains, la collecte d'un volume suffisant de données peut s'avérer extrêmement difficile. Les approches auto-supervisées permettent un gain de temps et de coûts, car elles remplacent tout ou partie des données de formation étiquetées manuellement.

Il existe trois approches différentes du NLP :

NLP basé sur des règles

Les premières applications de NLP étaient des decision trees simples, nécessitant des règles préprogrammées. Elles ne pouvaient fournir des réponses qu'à des prompts spécifiques, comme la version originale de Moviefone qui disposait de fonctionnalités rudimentaires de génération automatique de textes en langage naturel (NLG). Du fait que le NLP basé sur des règles ne possède pas de capacité de machine learning ou d'IA, cette fonction est très limitée et n'est pas évolutive.

NLP statistique

Développé plus tard, le NLP statistique permet d'extraire, de classer et d'étiqueter automatiquement les éléments des données textuelles et vocales, puis d'attribuer une probabilité statistique à chaque signification possible de ces éléments. Il repose sur le machine learning, permettant une analyse linguistique sophistiquée, comme le marquage des

parties du discours.

Le NLP statistique a introduit une technique clé : la mise en correspondance des éléments linguistiques, tels que les mots et les règles grammaticales, avec une représentation vectorielle, permettant de modéliser le langage à l'aide de méthodes mathématiques (statistiques), telles que la régression ou les modèles de Markov. Cette approche a inspiré les premiers développements du NLP, notamment les correcteurs d'orthographe et la saisie de texte T9 (texte sur 9 touches, utilisé sur les téléphones à clavier).

NLP par apprentissage profond

Plus récemment, les modèles d'apprentissage profond sont devenus le mode dominant du NLP, exploitant d'énormes volumes de données brutes et non structurées (textuelles et vocales) pour gagner en précision. L'apprentissage profond peut être considéré comme une évolution du NLP statistique, mais il utilise des modèles de réseaux neuronaux. Il existe plusieurs sous-catégories de modèles :

Modèles *séquence à séquence* (seq2seq) : basés sur des réseaux neuronaux récurrents (RNN), ils sont principalement utilisés pour la traduction automatique, en convertissant une phrase d'un domaine (comme l'allemand) en une phrase d'un autre domaine (comme l'anglais).

***Modèles transformateurs* :** ces modèles utilisent la tokénisation (la position de chaque token — mots ou sous-mots) et l'auto-attention (capturant les dépendances et les relations) pour calculer les relations entre les différentes parties du texte. Les modèles transformateurs peuvent être entraînés de manière efficace grâce à l'apprentissage auto-supervisé sur des bases de données textuelles massives. Un tournant dans le domaine des modèles transformateurs a été marqué par le modèle BERT (Bidirectional Encoder Representations from Transformers) de Google, qui est devenu et reste la base du fonctionnement de son moteur de recherche.

***Modèles autorégressifs* :** ce type de modèle transformateur est spécifiquement entraîné pour anticiper le mot suivant dans une séquence, représentant ainsi une avancée majeure dans la génération de texte. Parmi les exemples de LLM autorégressifs, on peut citer GPT, Llama, Claude et le logiciel open source Mistral.

***Modèles de fondation* :** les modèles de fondation préconstruits et conservés peuvent accélérer le lancement d'un projet NLP et renforcer la confiance dans son fonctionnement. Par exemple, les modèles de fondation d'IBM Granite sont largement applicables à divers secteurs d'activité. Ils prennent en charge des tâches NLP, notamment la génération de contenu et l'extraction d'informations. En outre, ils facilitent la génération renforcée par la recherche, un cadre permettant d'améliorer la qualité des réponses en reliant le modèle à des sources de connaissances externes. Ces modèles réalisent également la reconnaissance des

entités nommées, une tâche qui consiste à identifier et extraire des informations clés d'un texte.

Tâches de NLP

Le NLP décompose les données textuelles et vocales humaines en plusieurs tâches de manière à aider l'ordinateur à donner un sens à ce qu'il ingère. Ces tâches comprennent :

Résolution de coréférences

Reconnaissance des entités nommées

Marquage des parties du discours

Désambiguïsation du sens des mots

Résolution de coréférences

Cette tâche consiste à identifier si et quand deux mots renvoient à la même entité. L'exemple le plus courant consiste à déterminer la personne ou l'objet auquel un certain pronom fait référence (par exemple, « elle » = « Marie »), mais il peut également s'agir d'identifier une métaphore ou une expression idiomatique dans le texte (par exemple, quand un « ours » se réfère à une personne bourrue et non à un animal).

Reconnaissance des entités nommées (NER)

La reconnaissance des entités nommées (NER) identifie les mots ou les expressions en tant qu'entités utiles. Par exemple, elle reconnaît « Londres » comme un lieu ou « Maria » comme un nom de personne.

Marquage des parties du discours

Aussi appelé marquage grammatical, ce processus consiste à déterminer la catégorie grammaticale d'un mot ou d'un segment de texte en fonction de son utilisation et de son contexte. Par exemple, le marquage grammatical identifie « devoir » comme un verbe dans « Il va devoir partir en train » et comme un nom dans « Je dois faire mon devoir de maths ? »

Désambiguïsation du sens des mots

Ce processus sélectionne la bonne signification d'un mot qui en a plusieurs. Il s'appuie sur une analyse sémantique pour interpréter le mot dans son contexte. Par exemple, la désambiguïsation du sens des mots aide à distinguer le sens du verbe « faire » dans « faire ses preuves » (réussir) par rapport à « faire un gâteau » (cuisiner). Un système NLP sophistiqué est nécessaire pour comprendre des phrases complexes comme « une mûre mûre murmure au mur », où il faut distinguer entre des homonymes et des significations multiples.

Comment fonctionne le NLP ?

Le NLP combine diverses techniques informatiques pour analyser, comprendre et générer le langage humain de manière à ce que les machines puissent le traiter. Voici un aperçu d'un pipeline NLP typique et de ses étapes :

Prétraitement du texte

Le prétraitement du texte dans le cadre du NLP prépare le texte brut à l'analyse en le transformant dans un format plus facilement compréhensible par les machines. Le processus commence par la tokenisation, qui divise le texte en unités plus petites telles que des mots, des phrases ou des expressions. Cela permet de fragmenter un texte complexe en éléments plus gérables. Ensuite, une mise en minuscules est appliquée pour normaliser le texte, en convertissant tous les caractères en minuscules, afin que des mots comme « Pomme » et « pomme » soient traités de la même manière. Une autre étape courante est la suppression des mots vides, où les termes fréquemment utilisés, comme « est » ou « le », sont filtrés car ils n'apportent pas de signification notable au texte. La racinisation, ou la lemmatisation, réduit les mots à leur radical (par exemple, « partiront » devient « partir »), ce qui facilite l'analyse en regroupant différentes variantes d'un même mot. Le nettoyage du texte vient ensuite supprimer les éléments indésirables, comme la ponctuation, les caractères spéciaux et les chiffres, qui pourraient perturber l'analyse.

Une fois le prétraitement terminé, le texte est propre, normalisé et prêt à être interprété efficacement par les modèles de machine learning.

Extraction de caractéristiques

L'extraction des caractéristiques consiste à convertir le texte brut en représentations numériques que les machines peuvent analyser et interpréter. Cela implique de transformer le texte en données structurées à l'aide de techniques NLP telles que le sac de mots et le TF-IDF, qui quantifient la présence et l'importance des mots dans un document. Parmi les méthodes plus avancées figurent les plongements lexicaux comme Word2Vec ou GloVe, qui représentent les mots sous forme de vecteurs denses dans un espace continu, capturant ainsi les relations sémantiques entre les termes. Les plongements contextuels enrichissent encore cette représentation en tenant compte du contexte dans lequel les mots apparaissent, permettant ainsi d'obtenir des représentations plus nuancées et fidèles.

Analyse de texte

L’analyse de texte implique l’interprétation et l’extraction d’informations pertinentes à partir de données textuelles grâce à diverses techniques de calcul. Ce processus comprend des tâches telles que l’étiquetage de la classe de mots (POS, Part-of-Speech), qui identifie les rôles grammaticaux des mots, et la reconnaissance d’entités nommées (NER, Named Entity Recognition), qui détecte des entités spécifiques comme les noms, les lieux et les dates. L’analyse des dépendances examine les relations grammaticales entre les mots pour comprendre la structure des phrases, tandis que l’analyse des sentiments détermine le ton

émotionnel du texte, évaluant s'il est positif, négatif ou neutre. La modélisation thématique identifie les thèmes ou sujets sous-jacents dans un texte ou dans un corpus de documents. La compréhension du langage naturel (NLU, Natural Language Understanding) est un sous-ensemble du PNL qui se concentre sur l'analyse du sens des phrases. La NLU permet aux logiciels de trouver des significations similaires dans différentes phrases ou de traiter des mots ayant des significations différentes. Grâce à ces techniques, l'analyse de texte PNL transforme le texte non structuré en informations exploitables.

Entraînement du modèle

Les données ainsi traitées servent ensuite à entraîner des modèles de machine learning, qui apprennent à partir des modèles et des relations présents dans ces données. Pendant l'entraînement, le modèle ajuste ses paramètres pour minimiser les erreurs et améliorer ses résultats. Une fois entraîné, le modèle peut être utilisé pour faire des prédictions ou générer des résultats à partir de nouvelles données inédites. L'efficacité de la modélisation NLP est continuellement améliorée à travers des processus d'évaluation, de validation et de réglage fin afin d'améliorer la précision et la pertinence dans les applications concrètes.

Différents environnements logiciels sont utilisés tout au long de ces processus. Par exemple, la boîte à outils du langage naturel (NLTK) est une suite de bibliothèques et de programmes écrits en Python pour l'anglais. Elle prend en charge des tâches telles que la classification des textes, la tokenisation, le racinisation, le marquage, l'analyse syntaxique et le raisonnement sémantique. TensorFlow est une bibliothèque de logiciels gratuite et open source pour le machine learning et l'IA qui peut être utilisée pour entraîner des modèles pour les applications NLP. De nombreux tutoriels et certifications sont disponibles pour ceux qui souhaitent se familiariser avec ces outils.
Les défis du NLP

Même les modèles de NLP les plus avancés ne sont pas parfaits, tout comme la parole humaine est sujette aux erreurs. Comme toute technologie d'intelligence artificielle, la PNL comporte des pièges potentiels. Le langage humain regorge d'ambiguïtés, ce qui rend difficile pour les programmeurs de créer des logiciels capables de déterminer avec précision le sens de données textuelles ou vocales. Apprendre un langage humain peut prendre des années, et beaucoup continuent d'apprendre toute leur vie. Mais les programmeurs doivent enseigner aux applications de langage naturel à reconnaître et comprendre les irrégularités afin de garantir que ces applications soient précises et utiles.Les risques associés peuvent inclure :

Entraînement biaisé

Comme pour toute fonction d'IA, des données biaisées utilisées lors de l'entraînement 'du modèle fausseront les résultats. Plus les utilisateurs d'une fonction NLP sont diversifiés, plus ce risque devient significatif, notamment dans des domaines tels que les services

gouvernementaux, les soins de santé ou les ressources humaines. Les jeux de données des entraînements issus du web, par exemple, sont particulièrement exposés au biais.

Erreurs d'interprétation

Comme en programmation, il existe une règle d'or pour les modèles de NLP : « de mauvaises informations sont synonymes de mauvaises conclusions » La reconnaissance vocale, également appelée Speech to Text, consiste à convertir de manière fiable des données vocales en données textuelles. Cependant, les solutions NLP peuvent rencontrer des difficultés si les données vocales sont exprimées dans un dialecte obscur, marmonnées, truffées d'argot, d'homonymes, de grammaire incorrecte, d'expressions idiomatiques, de fragments, de fautes de prononciation, ou encore si elles sont enregistrées avec un bruit de fond important.

Nouveau vocabulaire

De nouveaux mots sont constamment inventés ou importés, et les conventions grammaticales peuvent évoluer ou être intentionnellement enfreintes. Dans ces cas, le NLP peut soit faire une estimation, soit admettre son incertitude, ce qui ajoute un niveau de complexité supplémentaire.

Ton de la voix

Lorsque les gens parlent, leur intonation ou même leur langage corporel peut transmettre un sens différent des mots seuls. L'exagération, l'accentuation de certains mots pour créer un effet ou le sarcasme peuvent être mal interprétés par le NLP, ce qui rend l'analyse sémantique plus complexe et parfois moins fiable.

Cas d'utilisation du NLP par secteur d'activité

Les applications de NLP se retrouvent désormais dans pratiquement tous les secteurs d'activité.

Finances:

Dans les transactions financières, des nanosecondes peuvent faire la différence entre succès et échec lorsqu'il s'agit d'accéder à des données ou d'effectuer des opérations. Le NLP peut accélérer l'extraction d'informations à partir de documents financiers, de rapports annuels et réglementaires, de communiqués de presse ou même de médias sociaux.

Soins de santé:

Les nouvelles découvertes et avancées médicales peuvent apparaître à un rythme plus rapide que celui auquel de nombreux professionnels de la santé peuvent s'adapter. Les outils basés sur le NLP et l'IA peuvent contribuer à accélérer l'analyse des dossiers médicaux et des études de recherche, permettant ainsi de prendre des décisions médicales plus éclairées et d'aider à la détection, voire à la prévention, de conditions médicales.

Assurance:
Le NLP peut analyser les demandes d'indemnisation pour identifier des schémas révélateurs de problèmes potentiels et détecter des inefficacités dans le traitement de ces demandes, conduisant ainsi à une optimisation accrue du traitement et des efforts des employés.

Juridique:
Presque toutes les affaires juridiques nécessitent l'examen de volumes considérables de documents, d'informations de fond et de précédents juridiques. Le NLP peut contribuer à l'automatisation de la recherche de preuves juridiques, en facilitant l'organisation des informations, en accélérant leur analyse, et en veillant à ce que tous les détails pertinents soient pris en compte.