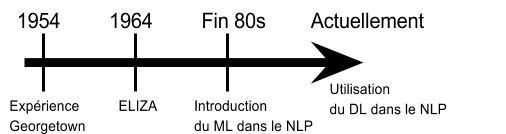
Introduction

Depuis les débuts de l’informatique, l’homme cherche à communiquer avec les machines. Si les nombreux langages de programmation permettent une forme d’échange entre l’homme et la machine, on aimerait que cette communication se fasse de façon plus naturelle. Pour que cela soit possible, il faut d’abord que la machine “comprenne” ce que l’utilisateur lui dit puis qu’elle soit capable de répondre d’une manière compréhensible par l’homme.

La discipline dernière ce processus s’appelle le Natural Language Processing (NLP) ou Traitement Automatique du Langage Naturel (TALN) en français. Elle étudie la compréhension, la manipulation et la génération du langage naturel par les machines. Par langage naturel, on entend le langage utilisé par les humains dans leur communication de tous les jours par opposition aux langages artificiels comme les langages de programmation ou les notations mathématiques.

Le NLP est parfois confondu avec le NLU dont l’expression est apparue plus récemment. Le NLU (Natural Language Understanding ou Compréhension du Langage Naturel en français) est en réalité un sous-domaine du NLP qui s'intéresse spécifiquement à la compréhension du langage écrit. Il regroupe des tâches comme l'analyse de sentiments , le résumé automatique de texte , les systèmes de questions-réponses , les agents conversationnels,....

Bref historique

[](https://images.ctfassets.net/95wnqgvmhlea/4ZCqHJRwwMGmuY0CSyCK8g/c7a98e21fe6fda33c807ad6dbc116e86/frise.png?fm=png)

Le NLP est un champ de recherche qui commence avec les débuts de l'informatique moderne et prend réellement son essor ces dernières années. Son histoire débute dans les années 40 et 50. Elle se concentre alors essentiellement sur la traduction de phrases simples. On peut citer l'expérience Georgetown en 1954 qui comportait la traduction complètement automatique de plus de soixante phrases russes en anglais.

Puis dans les années 60 et 70, on voit l'émergence des premiers chatbots comme ELIZA (1964). Mais c’est à la fin des années 80 que le NLP fait sa révolution avec l’introduction des algorithmes de Machine Learning dans le traitement du langage et l’augmentation de la puissance informatique.

Actuellement, avec les technologies informatiques toujours plus perfectionnées et abordables, la quantité de données open source toujours plus importante et l’utilisation du Deep Learning, le NLP est en pleine expansion. Le Deep Learning est utilisé dans de nombreuses tâches du NLP entraînant une amélioration significative de leur performance. Il est également à l’origine du développement d’une méthode de transformation de mots en vecteurs numériques appelée Word Embedding. Cette technique est une réelle avancée car elle permet d’obtenir des vecteurs quantitatifs qui conservent la similarité contextuelle des mots.

Quelques tâches du NLP

Le tâches du NLP sont nombreuses et plus ou moins complexes. Parmi les tâches courantes, on peut citer la segmentation de texte. Il en existe différents degrés. Pour la segmentation d’un texte en phrases, on parlera de Sentence Boundary Disambiguation (SBD). Il s’agit alors de déterminer où commence et où s’arrête une phrase. Pour cela, on peut s’appuyer sur un ensemble de règles pré-établies ou déterminées par apprentissage à partir d’un texte.

On peut aussi fractionner un texte en unités plus petites appelées tokens, on parle alors de tokenisation. Les tokens peuvent être des mots, des n-grammes (groupe de n tokens consécutifs), des chiffres, des symboles et de la ponctuation.

Enfin, il est possible de faire de la segmentation de topics. Plusieurs approches sont possibles :

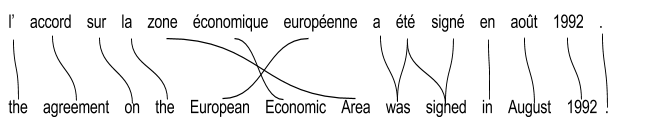
* Il peut s’agir d’une approche supervisée de classification de documents. Dans ce cas, les topics sont déjà connus et l’on entraîne un modèle de classification à partir d’un corpus de textes labellisés selon les topics pour ensuite utiliser ce modèle sur un nouvel ensemble de textes.
* Il peut également s’agir d’une approche non supervisée où l’on essaye de faire apparaître les topics principaux d’un ensemble de textes puis de relier les topics aux textes. On parle alors de Topic Modeling .

Les tâches de classifications ne se limitent pas aux topics. Il est possible de classer des textes ou bien des phrases selon leur polarité (positif versus négatif) par exemple. La classification est une approche dite supervisée. Elle nécessite d’avoir des jeux de données labellisés, ce qui n’est pas toujours aisé à trouver. On peut alors utiliser des approches non supervisées de clustering qui vont permettre de regrouper des phrases ou des textes similaires sans avoir besoin d’un jeu de données labellisé.

D’autres tâches très courantes en NLP sont le Part-Of-Speech Tagging (POS-Tagging ou étiquetage morpho-syntaxique en français) et le Named-Entity Recognition (NER ou reconnaissance d’entités nommées en français). La première consiste à associer à chaque mot d’un texte sa classe morphosyntaxique (nom, verbe, adjectif...) à partir de son contexte et de connaissances lexicales.

La deuxième tâche permet de reconnaître dans un texte un certain type de concepts catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates....

Enfin parmi les tâches de NLP plus complexes, on peut citer la traduction automatique (Machine Translation, MT en anglais). La traduction automatique statistique (Statistical Machine Translation, SMT en anglais) repose sur des algorithmes prédictifs qui “apprennent” à partir d’un corpus parallèle, c'est-à-dire un ensemble de textes en plusieurs langues, en relation de traduction mutuelle. La traduction automatique neurale (Neural Machine Translation, NMT en anglais) s’appuie sur des algorithmes de Deep Learning.

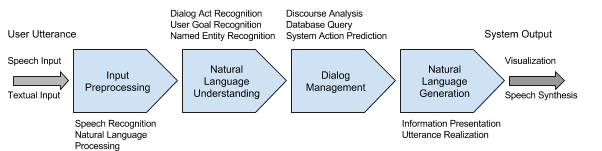


Toutes les tâches citées précédemment se concentrent sur l’analyse et la compréhension du langage naturel mais un autre aspect important du NLP est la génération automatique de textes (Natural Language Generation, NLG en anglais). Le NLG permet notamment d’automatiser la génération de rapports, de résumés, de paraphrases, de réponses dans un système de dialogue....

Exemples d’applications courantes en NLP

Toutes ces différentes tâches permettent de développer des applications et outils dont beaucoup nous servent au quotidien. Comme précisé précédemment, la traduction automatique est utilisée par Google Translate mais aussi par beaucoup d’autres outils de traduction comme l’application mobile Microsoft Translator permettant de traduire des documents dans une soixantaine de langues différentes ou encore Skype Translator pour faciliter les conversations de groupes allant jusqu’à 100 interlocuteurs.

On assiste à l’heure actuelle à une multiplication des chatbots et assistants personnels (Google Now, Cortana, Siri). Ces technologies font appel à plusieurs tâches de NLP pour “comprendre” la demande de l’utilisateur et y répondre sous forme de langage naturel.



Schématiquement, dans un processus de système de dialogue (en), la déclaration de l'utilisateur (sous forme vocale ou textuelle) est pré-traitée par diverses méthodes de NLP. Le résultat du pré-traitement est ensuite analysé par une unité de NLU. Puis les informations sémantiques sont analysées par le "dialog manager" qui garde l'historique et l'état du dialogue et gère le flux général du dialogue. La dernière phase va permettre de générer une réponse à l'utilisateur sous une forme comprise par celui-ci.

La classification de textes permet de construire des détecteurs de spams ou faire de l’analyse de sentiments pouvant être utilisée par exemple dans la prédiction d’une élection, du succès ou non d’un film ou tout autre produit.

Le clustering permet de regrouper des documents similaires et peut être utilisé dans un système de recommandation d’articles par exemple.

Enfin de nombreuses tâches de NLP sont utiles pour faire de l’extraction d’informations. Le Topic Modeling par exemple, permet de déterminer les thèmes principaux d’un document sans avoir à le lire et d’en extraire des mots clefs.