# LE TRAITEMENT AUTOMATIQUE DES LANGUES

## Comprendre les textes grâce à l'intelligence artificielle

**François-Régis Chaumartin**
Fondateur et CEO de Proxem

**Pirmin Lemberger**
Directeur scientifique chez onepoint

Préface d'**Olivier Delabroy**

DUNOD

Photo de couverture : © narvikk-iStock.





© Dunod, 2020
11 rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-080188-6

# Table des matières

# Préface

Qu'ont en commun la Grande marche du candidat Macron, le service client d'Auchan ou de Decathlon ? Le recrutement des cadres, la fiabilité des usines chez Air Liquide ou encore le chatbot du portail Orkyn ? Leur point commun c'est l'analyse sémantique de textes, réalisée par la startup Proxem.

Dans un monde où les données sont de plus en plus centrales et indispensables, nous nous retrouvons encore trop souvent dépassés par le tsunami qu'elles provoquent. Nos organisations ne peuvent plus ignorer le sujet et doivent de façon urgente se donner comme but de reprendre le contrôle de la donnée, grâce notamment à l'intelligence artificielle. Devenue un formidable vecteur de création de valeur avec les progrès spectaculaires de l'apprentissage automatique, l'IA peut en effet aider les entreprises à créer de la connaissance et des objets autonomes à partir de mégadonnées. Et leur permettre de rendre enfin possible un vieux rêve : donner la bonne information, à la bonne personne, au bon moment, au bon niveau/endroit de l'organisation.

Avec cet ouvrage, François-Régis Chaumartin, CEO de la startup française Proxem, et Pirmin Lemberger, directeur scientifique chez onepoint, nous embarquent dans une plongée au cœur de la linguistique et du *deep learning* ; ils nous font découvrir la puissance d'une branche spécifique de l'intelligence artificielle, le traitement automatique du langage naturel, c'est-à-dire des langues, écrites ou orales, parlées par les humains – en anglais *Natural Language Processing* (NLP). Ce livre s'adresse tout autant aux managers et cadres exécutifs des grandes entreprises – qui pourront explorer et prendre la mesure de tout le potentiel de création de valeur de cette branche de l'IA – qu'aux ingénieurs capables de faire précipiter cette valeur potentielle dans les comptes de l'entreprise, au travers de la réalisation de projets innovants. En fournissant au passage à ces deux populations de l'entreprise les clés pour se comprendre mutuellement.

La surcharge d'information textuelle est, de fait, devenue une réalité quotidienne, que la source soit interne (mails, documents et présentations, CRM, contrats…) ou externe (web, réseaux sociaux…). Toute organisation, quelle que soit sa taille, regorge de ces corpus de textes qui sont comme autant de mines d'or inexploitées. Devenons des chercheurs d'or ! Et faisons émerger grâce au NLP ces pépites de création de valeur qui contribuent à la croissance et à la compétitivité des entreprises : écouter la voix des clients grâce à l'analyse sémantique des verbatims collectés sur l'ensemble des canaux de vente, identifier les signaux faibles pour prendre de meilleures décisions et ainsi augmenter la satisfaction des clients et leur expérience ; prendre le temps d'écouter la voix des collaborateurs pour leur proposer une meilleure expérience dans une logique de symétrie des attentions ; réduire le temps d'analyse des contrats ou d'appels d'offre, ce qui ne manquera pas de révolutionner le travail des juristes d'entreprise ; rendre les usines plus performantes en analysant tous les rapports d'incidents et renforcer encore la connaissance métier ; fidéliser un écosystème de patients en proposant un dialogue au travers d'un chatbot, etc.

Les auteurs nous font prendre conscience (non sans une certaine délectation) de toute la complexité de l'analyse sémantique, nous faisant redécouvrir la richesse de la langue, sa subtilité, les magnifiques ambiguïtés qu'elle peut et sait créer, l'importance d'une virgule (« *on mange, les enfants* »)… Mais loin de se cantonner aux grands enjeux théoriques, les deux auteurs ont également un objectif très concret à court terme : aider les professionnels à réussir leur projet d'IA incluant une dimension de traitement du langage naturel. L'enjeu est de taille pour les grands groupes et les acteurs économiques qui pour l'essentiel ne maîtrisent pas encore cette nouvelle compétence aux frontières des technologies numériques, des sciences cognitives, de la littérature et des sciences humaines et sociales. Oui, dans ce monde numérique, la maîtrise de l'intelligence artificielle est incontestablement une des clés du pouvoir de demain et reprendre le contrôle de la donnée est désormais une question de résilience et de durabilité pour nos entreprises.

Si je devais retenir seulement deux conseils sur comment mener avec succès la transformation digitale au sein d'une entreprise, ce seraient les suivants. Tout d'abord partir systématiquement des usages et de la valeur créée avant de plonger dans la complexité de la technologie et de la gestion de projet. Cette conviction forte est partagée par François-Régis et c'est donc sans surprise que les auteurs consacrent leur premier chapitre à la création de valeur. Une fois la valeur identifiée, il faut passer à l'action et monter le projet qui fera briller la pépite brute et matérialiser la valeur créée dans les comptes de l'entreprise. Deuxième conseil que je peux donner, fort de mes années passées à piloter la transformation digitale d'un grand groupe industriel : mettre en place une démarche inclusive, rassembler des compétences multidisciplinaires dépassant les sacro-saints organigrammes, oser faire travailler ensemble opérationnels, marketing, digital et IT dès le premier jour du projet, être obsédé par la gestion du changement et toujours mettre l'humain avant la technologie.

Sur ce sujet d'ailleurs une opportunité se dessine pour notre vieille Europe, entre la Chine et les GAFA. Dans cette course au leadership mondial, l'Europe a la chance de proposer une approche unique, mettant l'humain et l'éthique au cœur de sa vision sur l'intelligence artificielle. Automatiser les tâches qui peuvent l'être, bien sûr, mais sans pour autant remplacer l'humain, et en mettant au contraire l'IA et la technologie au service de l'homme, pour l'aider à prendre de meilleures décisions. Et non les prendre à sa place. Cette valeur qui m'a guidé depuis le premier jour dans la mise en œuvre de la transformation chez Air Liquide est également au cœur des convictions et de la vision des auteurs de ce livre utile.

Olivier Delabroy

*Vice-Président Digital Transformation d'Air Liquide*

# Avant-propos

## ◆ *Objectifs de l'ouvrage*

L'intelligence artificielle est le moteur d'une nouvelle révolution industrielle. Si depuis plus de cinquante ans périodes d'euphorie et de déception ont alterné, les percées récentes en apprentissage automatique (*machine learning* ou ML) sont spectaculaires. Forts de cette dynamique, les GAFA, BATX et NATU[1] promettent monts et merveilles ; leurs annonces suscitent des fantasmes qui oscillent entre le rêve d'un monde meilleur et le cauchemar d'une dictature numérique. Les promesses n'ont jamais été aussi fortes : vaincre le cancer, traduire une conversation téléphonique en temps réel, anticiper les crimes ou encore se faire conduire dans une voiture autonome sans avoir à tenir le volant.

Une attente particulière concerne la compréhension de la langue humaine, comme en témoigne les succès récents des agents conversationnels (*chatbots*) et enceintes connectées. Le langage est une caractéristique différenciante de l'intelligence humaine et notre mode d'échange le plus naturel. Le traitement du langage est la branche de l'intelligence artificielle qui vise à traiter le texte, écrit ou oral, notamment en « comprenant » les mots et les rapports subtils qu'ils entretiennent entre eux. Appliquées au tsunami de données textuelles que nous recevons quotidiennement, ces technologies permettent d'automatiser des processus faisant intervenir la langue, d'améliorer l'écoute d'un écosystème grâce à l'analyse d'opinions ou encore de créer de nouvelles connaissances en fouillant des masses de documents.

Dès aujourd'hui, les applications du traitement automatique des langues facilitent la transformation digitale des entreprises en les aidant à prendre de meilleures décisions et à agir plus efficacement. De nombreux retours d'expériences récents présenteront ici des cas d'usages clairs dans des domaines aussi variés que l'expérience client, les ressources humaines, la voix des citoyens, les risques industriels, la veille économique, l'assistance juridique… Ces applications, mises en œuvre dans de grandes entreprises mais aussi des PME, témoignent du potentiel de ces technologies pour améliorer concrètement la performance des organisations.

S'il vous arrive d'avoir du mal à comprendre les ordinateurs, dites-vous bien que la réciproque est vraie. « Parler humain » est loin d'être simple : il suffit pour s'en convaincre d'apprendre une langue étrangère. Les mécanismes sous-jacents à l'intelligence artificielle appliquée au traitement des langues restent complexes à décrypter, car mettant en œuvre des maths de haut niveau. Doivent-ils pour autant être réservés à une élite ? Cet ouvrage démystifie ces technologies d'une façon pédagogique,

---

1. Les géants du web américains sont Google, Apple, Facebook et Amazon (parfois nommés GAFAM si on y intègre Microsoft). Leurs équivalents chinois sont Baidu (moteur de recherche), Alibaba (e-commerce), Tencent (messagerie et réseau social) et Xiaomi (entreprise technologique). NATU désigne Netflix, Airbnb, Tesla et Uber.

pour permettre aux professionnels de comprendre les nouvelles possibilités offertes, sans confondre science et science-fiction. Enfin, il aborde aussi les questions éthiques, notamment le rôle de l'humain dans une société de plus en plus automatisée.

### ◆ *Des révolutions industrielles aux révolutions numériques*

La machine à vapeur est le symbole de la première révolution industrielle. Les travaux de la fin du XVIII^e^ siècle au Royaume-Uni rendent possible la transformation de l'énergie thermique de la vapeur d'eau en énergie mécanique. Pour la première fois, l'homme maîtrise une source d'énergie mécanique pouvant être actionnée à la demande. Cette machine à vapeur utilise du charbon, seul combustible capable à l'époque de fournir suffisamment de chaleur pour produire de la vapeur. De nouvelles sources d'énergie vont de pair avec de nouveaux matériaux : la fonte se répand et favorise la construction de ponts puis de chemins de fer parcourus par des locomotives à vapeur qui réduisent les temps de trajet. Avec l'invention du télégraphe électrique et du code Morse, l'information aussi circule beaucoup plus vite.

La deuxième révolution industrielle commence à la fin du XIX^e^ siècle. L'apparition de nouvelles sources d'énergie (pétrole et électricité) et de matériaux nouveaux (acier et aluminium) provoque l'essor des grandes industries. Au fil du XX^e^ siècle, le moteur électrique ou thermique remplace progressivement la machine à vapeur. L'automobile puis l'avion raccourcissent les distances. Radio et télévision rendent possible une communication de masse instantanée. Le téléphone permet aux individus de se parler quelle que soit la distance qui les sépare.

La troisième révolution industrielle apparaît entre 1970 et 2000. Une fois encore, ce moment coïncide avec l'émergence de nouvelles sources d'énergie (le nucléaire s'impose en tant qu'alternative au choc pétrolier) et des matériaux innovants (silicones, céramiques, résines). L'informatique devient une industrie planétaire et les algorithmes font des pas de géant. La puissance de calcul des ordinateurs augmente d'une façon réellement exponentielle ; le volume de données disponibles explose aussi, grâce au Web et plus récemment aux objets connectés.

Chaque révolution industrielle provoque des changements majeurs dans les formes de production et les sources d'énergie, avec de nouveaux marchés et de nouveaux biens de consommation. Elle a ses gagnants et ses perdants : les plus fortes valorisations boursières sont passées des usines mécanisées aux plateformes logicielles ; des emplois disparaissent, d'autres sont créés.

En économie, des tendances antagonistes d'affrontent. Le capitalisme confirme ses appétits pour les monopoles[2]. Mais dans le même temps, des colosses aux pieds d'argile peuvent sombrer en ayant à peine le temps de se rendre compte de leur chute. Des ruptures technologiques ont uberisé les précédents leaders de certains marchés[3].

---

2. Par exemple, Google a acquis plus de cent sociétés (YouTube, Waze, Nest, DeepMind…) et Disney a créé un géant des loisirs avec les rachats de Lucasfilm, Pixar, Fox et Marvel.

3. Kodak, Nokia, BlackBerry et Altavista en ont fait les frais, pour n'en citer que quelques-uns.

Les cartes sont rebattues de plus en plus vite. Nous n'avons pas complètement digéré le choc de la troisième révolution industrielle que la quatrième s'annonce déjà. Une combinaison inédite de technologies émergentes (intelligence artificielle, biotechnologies, imprimantes 3D, robots, drones et voitures autonomes…) aura des répercussions inimaginables sur notre futur quotidien. La moitié des emplois actuels seront impactés (et non détruits comme on le lit parfois).

L'intelligence artificielle est le symbole de cette quatrième révolution industrielle, comme la machine à vapeur fut celui de la première et le moteur à combustion celui de la deuxième. Mais au lieu de produire de l'énergie mécanique grâce au charbon ou au pétrole, l'IA aide à créer de la connaissance et des objets autonomes à partir de mégadonnées ; dire que les données sont le pétrole du XXIe siècle semble une analogie correcte.

Les humains ont utilisé des bêtes de trait pendant des milliers d'années pour les pénibles travaux agricoles. Machines à vapeur et moteurs ont fourni un substitut bien plus efficace de leurs muscles puis ouvert la voie à des usages nouveaux, inimaginables auparavant. Le même saut quantique se retrouve dans le passage de l'informatique classique à l'intelligence artificielle. Cette dernière rentre tout simplement en concurrence avec le cerveau humain, et le dépasse même désormais sur un nombre croissant de tâches. La révolution numérique en cours aura donc un impact encore plus colossal que les précédentes. Et si la première révolution industrielle a duré un siècle, la révolution numérique actuelle risque de changer la société en profondeur en une ou deux décennies.

### ◆ *Le langage, support de l'intelligence*

#### Les intelligences humaines

La seule intelligence réelle – l'**intelligence humaine** – reste un mécanisme mystérieux, difficile à comprendre ou à modéliser même au sein des meilleurs labos de recherche. Cette notion est sujette à de multiples interprétations, et nous ne nous risquerons pas à la définir formellement. De multiples manifestations de l'intelligence cohabitent[4], faisant appel à des capacités cognitives complémentaires : résoudre des problèmes abstraits, logiques et mathématiques ; comprendre les mots et leurs subtiles nuances de sens ; retrouver son chemin, en créant une image mentale de son environnement ; comprendre ses propres émotions et savoir se contrôler ; comprendre les autres et communiquer avec eux avec empathie ; exercer une coordination neuromusculaire fine ; jouer d'un instrument de musique ; classer correctement des objets en identifiant leurs points communs et leurs différences ; raisonner par analogie ; s'interroger sur le sens des choses… Ces capacités peuvent aussi être différenciées en intelligences **analytique** (résoudre des problèmes, analyser des résultats), **créative** (trouver une solution à un nouveau problème) ou **pratique** (adaptation à l'environnement).

4. Voir la page Wikipédia sur la Théorie des intelligences multiples.

Que ce soit pour discuter, avoir un œil critique sur une feuille Excel, écrire un poème ou conduire une voiture, nous pouvons compter sur notre cerveau d'humain qui est le produit de millions d'années d'évolution. À titre indicatif, il compte autour de 100 milliards de neurones (les cellules nerveuses) et un million de milliards de synapses (les connexions reliant les neurones)[5]. Le regain d'intérêt actuel pour l'intelligence artificielle est d'ailleurs lié aux résultats remarquables de l'apprentissage profond (*deep learning* ou DL), qui utilise des réseaux de neurones artificiels s'inspirant de leur cousin biologique.

La vision est la plus ancienne manifestation de l'intelligence : elle nous permet de reconnaître des formes et des mouvements. Depuis notre création, pour échapper aux prédateurs, nous avons appris à être très réactifs aux stimuli visuels. L'œil contient 70 % de nos capteurs sensoriels ; quand une image parvient à notre rétine, les formes, couleurs et autres informations sont analysées en parallèle par différentes zones du cerveau. La compréhension des mots impose en revanche à notre cerveau un traitement linéaire, qui exige plus de temps et d'énergie. Notre cerveau traiterait une image soixante-mille fois plus vite qu'un texte ; comme dit le proverbe attribué à Confucius, « une image vaut mille mots ».

L'évolution vers *homo sapiens* amorcée il y a cent mille ans nous a ensuite fait don du langage et de la capacité à représenter des abstractions, y compris le mensonge. L'apparition de l'écriture il y a cinq mille ans était liée à la nécessité de conserver la trace des transactions, histoires et discussions. La fabrication des langues humaines a suivi un processus anarchique. De très nombreux peuples ont défini leur propre langue, en formalisant au cours des siècles leur lexique et leur syntaxe. La rencontre avec d'autres populations au fil du temps a créé des échanges, des incompréhensions, des assimilations, des partages[6] et des métissages qui continuent de nos jours : on parle parfois « franglais » chez nous, « denglish » en Allemagne et « portugnol » à la frontière entre le Brésil (dont la langue officielle est le portugais) et d'autres pays d'Amérique du Sud (qui parlent espagnol). Cette joyeuse cacophonie est rentrée dans l'histoire avec le mythe de la tour de Babel ; elle représente aussi un défi pour les modules de détection de langue.

On qualifie de langues naturelles celles parlées par les humains, par opposition aux langues construites ou informatiques. Du fait de leur historique, les langues naturelles sont des artefacts complexes ; si elles s'appuient sur des règles, les exceptions y sont nombreuses et connaissent elles-mêmes des exceptions[7]. Est-ce que le français est particulièrement complexe ? Non. Le russe paraîtra compliqué à un

5. Un réseau de neurones artificiel, que nous décrirons au chapitre 5, en compte typiquement quelques dizaines de milliers. Le système DeepMind de Google, qui a représenté un investissement financier colossal, en compterait 10 milliards.

6. Rappelons que le mot anglais *flirt* vient du français « conter fleurette ».

7. Par exemple, en français, les verbes du 1er groupe finissent par un « -es » à la 2ème personne du singulier au présent de l'indicatif (« tu manges »). À l'impératif, les formes se terminant par un « e » muet ne prennent toutefois pas de « s » (« mange »)… sauf s'ils sont immédiatement suivis des pronoms « en » ou « y » (« manges-en »). Cette règle vaut pour le pronom « en » et non pour la préposition, car on écrit bien : « mange en silence ».

locuteur français à cause de l'alphabet cyrillique, et le chinois encore plus du fait des sinogrammes. Ceux qui ont la chance d'être polyglotte disent qu'il n'existe pas vraiment de langue naturelle plus complexe ou plus simple qu'une autre[8]. Les difficultés que nous, humains, rencontrons en découvrant une langue étrangère ne sont d'ailleurs rien comparées à l'effort nécessaire pour enseigner une langue humaine à une machine qui ne sait fondamentalement que manipuler des 0 et des 1.

Les mécanismes du cerveau et des réseaux de neurones biologiques semblent donc mieux adaptés au traitement des images qu'à celui du texte, dont les règles sont souvent arbitraires. S'il s'avère que leurs cousins artificiels utilisés en apprentissage automatique partagent cette caractéristique, cela fournirait une explication au fait que la capacité de l'IA pour comprendre réellement du texte reste modeste, en tout cas en retrait des performances exceptionnelles constatées en reconnaissance d'images.

### Les intelligences artificielles

L'**intelligence artificielle** peut être définie comme l'ensemble des théories, technologies et techniques permettant à la machine de simuler l'intelligence humaine.

On a cru pendant longtemps que la machine ne pourrait pas battre le meilleur joueur humain dans des jeux de stratégie complexes. Mais le programmes d'IBM (*Deep Blue*) a battu le champion du monde des échecs Garry Kasparov en 1997. Les humains seraient-ils mauvais joueurs ? Beaucoup ont jasé à l'époque en prétendant que, finalement, bien jouer aux échecs ne nécessite pas tant d'intelligence que de puissance de calcul. Un autre jeu de stratégie – le go – dont la combinatoire est encore plus grande qu'aux échecs, devenait le nouveau symbole de la résistance des (meilleurs joueurs) humains face aux machines. Las, en 2016, le programme de la filiale DeepMind de Google (*AlphaGo*) l'a emporté face au champion du monde Lee Sedol. Dans les deux cas, ce succès a été rendu possible par des approches technologiques disruptives, une énorme puissance de calcul et des budgets pharaoniques[9]. « Intelligence artificielle » est un mot-valise dont les contours varient avec le temps : spécifier ce qui en relève ou non est un débat – souvent stérile – dans lequel nous ne rentrerons pas ici.

La conduite d'une voiture sur route libre est pour l'instant réservée aux humains qui ont passé avec succès le permis. Il est possible que des voitures autonomes, mues par un système informatique avancé, apparaissent sur nos routes dans quelques mois. Elles sont déjà autorisées à titre expérimental dans quelques régions du monde (en Californie, par exemple). Ce projet fait l'objet d'investissements de plusieurs milliards de dollars de la part de géants de la data (GAFA, Uber…) mais aussi des principaux constructeurs automobiles (Toyota, Ford, Mercedes, Renault…). Les assureurs y voient l'aubaine d'une diminution des sinistres – et des remboursements.

---

8. L'espéranto est une langue simple, mais il s'agit justement d'une langue artificielle conçue dans le but d'être facile à apprendre.

9. Le rachat de DeepMind par Google a été de l'ordre du demi-milliard de dollars. Son programme AlphaGo Zero a écrasé son prédécesseur AlphaGo (qui avait battu le meilleur joueur de go) en apprenant à jouer contre lui-même plusieurs millions de parties. Le coût de la puissance de calcul nécessaire pour reproduire l'apprentissage a été estimée à environ 30 millions de dollars.

Toutes les IA ne sont pas égales entre elles. La science-fiction a abondamment exploité l'idée des machines pensantes, susceptibles de se révolter contre leurs créateurs (Terminator ou HAL dans *2001, l'Odyssée de l'espace*) ou au contraire d'aider à rendre le monde meilleur (dans le film *Her*). Une telle super IA serait capable de penser par elle-même et dotée d'une conscience de soi. Qualifiée d'**IA forte** ou d'**IA générale**, elle pourrait – entre autres exploits – passer avec succès le test de Turing, c'est-à-dire échanger des messages textuels avec un humain qui ne s'apercevrait même pas que son interlocuteur n'en est pas un.

Commençons par tordre le cou à une idée romantique : une telle IA forte relèvera du domaine de la SF pour encore longtemps. La théorie de l'apprentissage elle-même limite l'apparition d'une IA générale. Les travaux de Vapnik et Chervonenkis nous enseignent qu'aucun algorithme d'apprentissage ne peut bien fonctionner d'une manière transverse sur des problématiques différentes. Dit autrement, un algorithme qui fonctionne efficacement sur un problème particulier sera moins performant en moyenne sur les autres problèmes. Le folklore mathématique appelle joliment ceci le *"No Free Lunch Theorem"* (« pas de déjeuner gratuit »). Le monde est trop complexe pour qu'il existe un algorithme général permettant de tout apprendre sans faire l'objet d'une adaptation au problème particulier à traiter : du sur-mesure s'impose souvent.

Pour l'instant, une intelligence artificielle ne sait apprendre à résoudre qu'une tâche donnée. Mais qu'est-ce que cela veut dire en pratique ? La machine va, par exemple, savoir apprendre à reconnaître des concepts tels que les marques, produits ou opinions dans les textes écrits par des consommateurs. Ou encore reconnaître un type particulier de formes dans des photos, comme des chatons, des voitures ou des quiches. Ou savoir identifier des personnes dans des vidéos et lire sur leurs lèvres. En 2020, sur l'ensemble de ces exemples, et beaucoup d'autres, la machine obtient des performances supérieures à celles des humains : on parle d'ailleurs alors de performances surhumaines.

### L'IA pour comprendre le langage humain

Le **traitement automatique des langues** (*natural language processing* ou *NLP*)[10] est la branche de l'intelligence artificielle qui vise à traiter les langues parlées par les humains. Cela couvre un grand nombre d'applications, comme retranscrire de l'oral en écrit, générer automatiquement une synthèse, traduire un document, chercher de l'information ou encore « comprendre » ce que veut dire un texte. Dans ce dernier cas, on emploie aussi souvent les termes voisins d'**analyse sémantique** (*semantic analysis*) et de **fouille de textes** (*text mining*).

La spécificité de ces disciplines scientifiques est de cumuler un grand nombre de tâches complexes et de problèmes non résolus à ce jour : résolution d'anaphores, désambiguïsation lexicale, analyse syntaxique, correction orthographique, prise en compte des figures de styles… Cette complexité résulte des nombreuses ambiguïtés présentes dans les langues naturelles. Comprendre le sens d'un texte nécessite

10. On emploie parfois aussi l'expression linguistique informatique (*computational linguistics*), notamment pour des travaux théoriques.

la résolution de problèmes unitaires, relevant souvent de la recherche fondamentale, à l'intersection de plusieurs disciplines : informatique théorique, mathématique, linguistique, psychologie cognitive, enseignement des langues…

Comme d'autres branches de l'intelligence artificielle, la compréhension de textes est un domaine qui a périodiquement soulevé de grands espoirs, avant qu'ils ne retombent, les réalisations n'étant pas à la hauteur des attentes. La traduction automatique en est une bonne illustration. La tâche est notoirement complexe ; les Italiens ne disent-ils pas *traduttore, traditore* (« traduire, c'est trahir ») ? Après la seconde guerre mondiale, des budgets militaires conséquents ont été investis dans cette discipline par les Américains et les Russes qui voulaient comprendre ce qui était dit par l'autre camp. Las, le rapport Bar-Hillel concluait en 1960 à l'impossibilité d'une traduction automatique de qualité humaine, gelant tout investissement dans le domaine pendant des années. Les années 1980 ont vu apparaître de nouvelles approches et le domaine a progressé, avant de stagner à nouveau. Le retour des réseaux de neurones a permis ces trois dernières années des progrès spectaculaires grâce aux efforts des GAFA, de Systran ou DeepL. À défaut d'être parfaits, les résultats se sont tellement améliorés ces derniers mois qu'il devient impossible de détecter à l'œil nu – même pour un œil exercé – qu'un document est issu d'une traduction automatique.

Même si on reste actuellement loin d'une compréhension en profondeur comme l'effectue un humain, extraire automatiquement du sens d'un texte devient un objectif réaliste. En effet, chaque grande vague d'intelligence artificielle a apporté son lot de progrès. Des algorithmes nouveaux ont permis de progresser sur la plupart des tâches.

Les techniques d'apprentissage automatique ont montré leur efficacité, en remplacement ou en complément des systèmes experts à base de règles ; leur essor a été rendu possible par l'apparition d'un nombre croissant de corpus[11] annotés manuellement, permettant cet apprentissage. De nouveaux algorithmes sont apparus, dispensant même de l'étape d'annotation manuelle. La puissance de traitement des machines et leurs capacités de stockage doublent régulièrement. Nous sommes donc en présence d'une conjonction de facteurs favorables aux progrès dans ce domaine.

### ◆ *Plan de l'ouvrage*

Le premier chapitre présente les **applications et cas d'usage** du traitement du langage. Il survole les familles de technologies pour donner une idée des possibilités offertes, puis détaille comment ces applications peuvent concrètement servir aux différentes directions d'une organisation.

Le chapitre 2 rappelle les **bases de la linguistique** et les différentes formes d'ambiguïté présentes dans les langues naturelles. Il présente ensuite les spécificités de la linguistique informatique et quelques-uns des enjeux rencontrés lors de l'analyse sémantique au niveau des différents constituants du texte.

---

11. Un corpus est un ensemble de textes cohérents, représentatif d'un certain usage de la langue.

L'enjeu du NLP est de transformer du texte en une représentation formelle permettant d'effectuer un calcul. La question complexe de la **représentation du sens** d'un énoncé se pose donc. Nous verrons dans le chapitre 3 comment la définir. Puis, une fois que des informations et connaissances sont extraites du texte, comment les lier à des bases de connaissances externes (thésaurus, taxonomies, ontologie…).

Le chapitre 4 présente les **principales tâches de NLP** et les difficultés à surmonter : prétraitements (détection de la langue, segmentation, correction d'orthographe, etc.), analyse syntaxique, analyse sémantique (reconnaissance d'entités, extraction de relations, résolution d'anaphores, analyse d'opinions…). Nous verrons aussi les tâches de génération de texte et d'extraction terminologique.

Le chapitre 5 introduit les technologies de **machine learning** et de **deep learning** appliquées au traitement du langage. Réseaux de neurones, schéma encodeur-décodeur et mécanisme d'attention n'auront plus de secret pour le lecteur. Nous détaillerons aussi comment les modèles récents de deep learning (ULMFiT, ELMo, BERT et XLNet) améliorent toutes les tâches de NLP.

Ces percées technologiques permettent de proposer une **conversation artificielle** qui dépasse enfin le simple gadget. Le chapitre 6 aborde les architectures permettant de mettre en place un chatbot efficace. Il en approfondira deux, la première cherchant à limiter l'effort d'apprentissage, la seconde permettant d'injecter des règles métiers.

Comment **mener avec succès un projet** d'analyse sémantique dans la « vraie vie » ? Le chapitre 7 présente une approche méthodologique sur différents types de projets ; il détaille les différentes étapes de la démarche, de l'intégration d'un corpus jusqu'à l'obtention de connaissances exploitables. Une analyse du Grand débat national (sur l'axe de la transition écologique) permet d'illustrer concrètement cette démarche.

Enfin, le chapitre 8 conclura cet ouvrage en évoquant les **perspectives et problèmes ouverts**, qu'ils soient d'ordre éthique ou technologique.

### ◆ *À qui ce livre est-il destiné ?*

#### À l'utilisateur métier

Cet ouvrage est destiné en premier aux non-spécialistes de l'intelligence artificielle, qui brassent au quotidien des textes dans le cadre de leur activité professionnelle, et dont le besoin métier est de comprendre ce qui est exprimé dans des millions de documents ou d'automatiser des processus. Les experts techniques à qui ils expriment leur besoin métier leur répondent souvent dans un jargon technique, ce qui rend délicat un dialogue constructif. Qu'ils travaillent au sein d'une direction commerciale, marketing, ressources humaines, juridique, qualité ou innovation, les managers trouveront ici des retours d'expérience sur les cas d'usages qui les concernent directement. Ils auront aussi un cadre méthodologique et une explication pédagogique des technologies utilisées et de leurs limites actuelles : les notions de *machine learning* et d'apprentissage par transfert deviendront beaucoup plus claires.

Cet ouvrage aidera aussi les utilisateurs avertis travaillant dans une direction métier à devenir *citizen data scientist*[12]. Aujourd'hui, ce rôle devient complémentaire de celui du *data scientist* : sans le remplacer, il met à profit une expertise métier pour effectuer des tâches d'analyse d'une sophistication allant de simple à modérée. Un *citizen data scientist* peut ainsi se focaliser sur la création de modèles capables de résoudre ses problématiques métiers et d'améliorer la compétitivité de son organisation. Devant la pénurie de profils qualifiés en science des données, ce nouveau rôle est probablement appelé à prendre de l'importance.

### À l'ingénieur ou au *data scientist*

Cet ouvrage sera aussi précieux pour les ingénieurs, informaticiens ou *data scientists*, souhaitant analyser des textes pour en extraire des données. Il leur présentera un large panorama des cas d'usages et applications concrètes possibles.

Du fait d'une méconnaissance des problèmes linguistiques, la difficulté du traitement automatique des langues est souvent sous-estimée. Disposer de librairies *open source* en téléchargement gratuit, des derniers articles scientifiques des GAFA décrivant des algorithmes de deep learning, ou même de modèles de langues préentraînés n'a jamais suffi à couvrir directement un besoin métier dans un projet d'analyse sémantique. La raison en est simple : il faut pour chaque projet une adaptation à la langue, au domaine et aux spécificités du corpus à traiter.

L'ouvrage permet ainsi au lecteur de profil technique de comprendre l'intégralité des tâches de traitement du langage ainsi que la maturité des briques techniques actuelles. En acquérant ces nouvelles compétences, il peut devenir *text scientist* : les notions de désambiguïsation lexicale ou d'anaphore n'auront plus de secret pour lui.

### ◆ *Travaux pratiques*

À plusieurs reprises dans cet ouvrage, le logiciel Proxem Studio est utilisé afin d'illustrer le contenu et de le rendre plus concret. Cette plateforme, développée par la startup française Proxem, fournit un environnement complet et intégré pour la réalisation de projets de traitement automatique des langues.

Vous pouvez retrouver une version de démonstration du logiciel à l'adresse suivante : www.proxem.com/livre-nlp.

### ◆ *Remerciements*

Les auteurs souhaitent ici témoigner leur reconnaissance à ceux qui ont contribué à cet ouvrage par leur disponibilité ou les précieux partages de connaissances.

La section consacrée à l'état de l'art en matière d'approche statistique du NLP du chapitre 5 a bénéficié de manière déterminante de l'expertise de Thomas

12. Imaginée en 2016 par le cabinet de conseil en technologies Gartner, cette fonction allie expertise métier et *data science* pour extraire de la valeur des données afin de répondre aux besoins des métiers. Cette évolution est rendue possible grâce à la simplification des logiciels dédiés et à l'automatisation grandissante des tâches de *data science* et de *text mining*. Un manager devient ainsi capable de créer ou de générer en autonomie des modèles d'analyse avancés, sans être forcément un expert des probabilités ou de l'IA.

Scialom (chercheur en IA chez ReciTAL.ai, expert en deep learning appliqué au NLP). Nous remercions aussi pour leur contribution experte : Hugues de Mazancourt (VP innovation chez Yseop) sur la génération de textes ; Louise El Yafi (responsable éditoriale chez Doctrine) sur le NLP au service des avocats ; Bruno Soubiès (fondateur de disRHupt et auteur de *L'entreprise à l'écoute de son personnel* aux éditions Kawa) sur la partie RH ; Ariane Nabeth-Halber (directrice speech solutions chez Bertin IT) sur la reconnaissance vocale. Et les text scientists, ingénieurs et spécialistes du deep learning chez Proxem, notamment Thomas Perrais, José Coch, Laurie Marche-Lefèvre, Grégoire Telmon, Elisa Piccinini, Vincent Jacquelinet, Julien Perrin, Kira Kiranova, Francois Brown de Colstoun, Rachel Drozdz, Cécile Potier, Victor Camara et Hamada Saleh. Merci aussi à nos relecteurs.

Un clin d'œil amical des auteurs à Jean-Baptiste Mestelan pour la mise en relation.

Je tiens ici à remercier Olivier Reisse, partner chez onepoint et fondateur de weave Business Technology, pour ses encouragements et pour avoir su créer un environnement de travail à la fois exigeant et bienveillant dont cet ouvrage a largement bénéficié. Mes remerciements vont aussi à Gontran Peubez, partner onepoint, pour son soutien sans faille à notre activité de veille et de R&D sur l'IA.

*Pirmin Lemberger*

Je remercie chaleureusement tous les talents de la formidable équipe Proxem, sans qui nous n'aurions pu réussir autant de beaux projets, notamment mes associés : Jocelyn Coulmance, Nicolas Frelat, Éric Vernet, Thomas Cohu, Alain Garnier et *last but not least* Carole Timelli. Sans pouvoir les citer tous, merci aussi à nos clients qui ont accepté de témoigner publiquement sur les projets que nous avons réalisés ensemble depuis dix ans, notamment Thierry Roche (Apec), Jérôme Desreumaux (Auchan, Orkyn), Laure Sanchez (BlueLink), Kamal Harfaoui (Bouygues Telecom), François-Xavier Guérin (Carrefour), Hervé Andorre (Dassault Systèmes), Kocélla Mechouek (Decathlon), Jean-Rémy Dudragne (Engie), Xavier Fontana (Thales), Claude Fauconnet (Total), Guillaume Jourdan (Vitabella) ; sans oublier, chez Air Liquide, Thanos Kontopoulos, Jean André, Habiboulaye Amadou-Boubacar, Denise Mery… et évidemment Olivier Delabroy qui a aussi préfacé cet ouvrage. Merci à Sylvain Kahane et Laurence Danlos de m'avoir patiemment guidé pendant ma thèse.

Ce livre est dédié à mes parents, qui m'ont donné le goût des mots tout en me poussant vers les sciences, ainsi qu'à mes filles – Cerise, Emilie et Mahaut – à qui j'espère avoir transmis à mon tour cette double passion.

*François-Régis Chaumartin*


# Les applications et usages du NLP

Objectifs

**Ce chapitre présente différentes applications du traitement du langage. Il survole les familles de technologies pour donner au lecteur une idée des possibilités offertes, sans rentrer dans les détails techniques, et rappelle les fondamentaux des systèmes d'analyse sémantique, fouille de texte, analyse d'opinions, systèmes de recommandation, génération de textes et assistants virtuels (chatbots…). Il détaille ensuite comment ces applications peuvent être concrètement mises au service des différents métiers d'une entreprise : commerce et marketing, ressources humaines, services juridiques, mais aussi qualité, communication et innovation bénéficient des apports de ces solutions. On verra aussi comment le monde politique peut utiliser l'analyse sémantique, comme on l'a constaté avec un certain succès lors des élections présidentielles françaises en 2017.**

## 1. 1 PANORAMA DES TYPES D'APPLICATION

Imaginez que l'une de vos missions quotidiennes nécessite de lire des documents – pour les analyser, y répondre ou en tirer une synthèse – mais avec une volumétrie de texte telle que cela en devient impossible. Vous seriez ravi de disposer d'un assistant spécialisé pour vous aider à traiter chaque jour des milliers de documents importants pour votre organisation. Lui-même disposerait de plusieurs stagiaires ; chacun aurait une compétence certes limitée, car ne sachant réaliser qu'un type de tâche, mais à la vitesse de l'éclair.

Prenons comme exemple l'analyse des courriels reçus par le service client. Certains documents sont mal écrits ? Pas de souci, un stagiaire les corrigera avant l'analyse. D'autres ne sont pas dans la langue attendue ? Un autre stagiaire les traduira. L'un de stagiaires va donner des millions de coups de Stabilo dans ces documents pour surligner les éléments qui vous intéressent ; ce faisant, il constitue à partir des textes

une base de données d'informations structurées. Un autre stagiaire encore va classer les documents en fonction de leur thématique et de leur urgence. Un stagiaire différent enverra automatiquement une réponse personnalisée pour indiquer que la demande est prise en charge. Un autre va utiliser des algorithmes de *data mining* sur les informations extraites du texte pour faire une analyse prédictive (calcul de séries temporelles permettant de détecter des tendances dans le temps). Un nouveau stagiaire va, grâce au calcul de corrélations, détecter des signaux faibles qui peuvent être exploités pour améliorer la performance de l'organisation.

Vous l'aurez compris, notre « assistant » est virtuel – il s'agit d'une application – et ses stagiaires sont des tâches de traitement du langage. Comme des briques de Lego qui s'emboîtent, ces tâches peuvent être assemblées pour créer de nouvelles applications et paramétrées pour traiter un corpus particulier en tenant compte de ses spécificités. Comme nous le verrons dans le chapitre 4, certaines de ces tâches sont de bas niveau (comme la détection de langue d'un texte ou le découpage d'un texte en mots) : prises séparément, leur création de valeur est faible du point de vue d'un utilisateur métier. Au contraire, d'autres tâches de haut niveau (comme la traduction automatique, la correction orthographique, le résumé de texte ou la détection de plagiat) peuvent être considérées comme des applications à part entière.

Pour finir sur l'analogie entre une armée de stagiaires humains et un logiciel, ce dernier n'a évidemment pas besoin de se reposer ou de dormir et il analyse les milliers de documents beaucoup plus vite. Mais un autre intérêt existe à l'approche automatisée : la constance de l'analyse effectuée. Même si les systèmes de NLP ne sont pas encore tout à fait au niveau de l'humain en termes de compréhension du langage, ils connaissent moins de biais et de subjectivité. On se rend compte à l'usage que des humains ne portent pas toujours exactement le même jugement en lisant un texte. Pire, une même personne pourra émettre deux avis différents, selon son humeur ou sa fatigue du moment.

Nous présenterons ici une double typologie des applications de traitement du langage. Nous commencerons par aborder les familles de technologies mises en œuvre : extraction d'informations, analyse d'opinion, génération de texte, moteurs de questions/réponses, traduction automatique et agents conversationnels… en expliquant ce que ces applications peuvent réaliser (et non comment elles le font, ce qui relèvera des chapitres suivants).

Nous illustrerons ensuite concrètement comment de telles applications peuvent servir au quotidien au commerce, au marketing, aux ressources humaines, au juridique, à l'innovation… Toutes les directions d'une entreprise sont potentiellement bénéficiaires des applications de traitement du langage. Certaines sont fortement liées à un métier donné : par exemple, si le *matching* entre offres et CV relève des ressources humaines, l'aide à la lecture des contrats concerne plutôt les métiers juridiques. D'autres sont spécifiques à un secteur, notamment pour des raisons réglementaires : la santé, la finance ou la sécurité nationale, pour ne citer qu'elles, imposent des contraintes particulières. En revanche, une application de connaissance clients analysant les retours de consommateurs intéresse surtout les directions

commerce et marketing, mais concerne tous les secteurs B2C, de la grande distribution à la finance en passant par le transport aérien.

Les retours d'expériences présentés ici sont pour la plupart issus de projets réalisés ces dernières années par la société Proxem. Nous laisserons ponctuellement la parole aux experts travaillant chez d'autres éditeurs de logiciels spécialisés ou aux clients finaux pour qu'ils livrent leurs retours d'expériences.

## 1. 2 APPLICATIONS PAR FAMILLE DE TECHNOLOGIE

### 1.2.1 Analyse sémantique : comprendre le sens d'un texte

Que veut dire comprendre un texte ? Adeline Nazarenko propose la formulation suivante [1] : « *De manière abstraite, on peut considérer que "comprendre un texte" signifie être capable de modifier sa représentation du monde en fonction des informations véhiculées par le texte. Cela suppose qu'un être humain ou un système intelligent dispose d'un ensemble de connaissances qui constitue sa vision de son environnement physique, intellectuel, social et symbolique. Dans cette perspective, la compréhension se traduit par l'ajout, la suppression ou la correction de connaissances. En pratique, le niveau de compréhension dépend de l'objectif visé et de la nature du texte considéré. On ne lit pas un texte de loi ou une police d'assurance comme un article de presse, un manuel scolaire comme une notice pharmaceutique. En soi, la compréhension n'est pas une tâche. C'est une activité préalable à de nombreuses tâches, comme le résumé, la traduction, l'exécution d'instructions…* »

L'**analyse sémantique** est la branche du traitement automatique des langues qui vise à « comprendre » le sens d'un texte. Les guillemets sont de rigueur ici, car la représentation du sens qu'une machine est actuellement capable de produire est bien moins riche que celle qu'un humain aura en lisant un texte. Les langues humaines présentent en effet un grand nombre d'ambiguïtés qui en rendent la compréhension en profondeur complexe pour la machine ; et la compréhension d'un énoncé nécessite souvent de disposer de connaissances du monde, en plus de la simple lecture du texte, pour interpréter ce qui est explicitement écrit.

L'analyse sémantique repose sur la combinaison de plusieurs tâches d'informatique linguistique. L'état de l'art actuel de ces applications revient à extraire des informations du texte sous forme d'un graphe[1] de connaissances. Cette extraction d'informations permet de créer un nouveau canal de données qui vient compléter celles, déjà structurées, disponibles.

D'un point de vue quantitatif, transformer le texte en data grâce à l'analyse sémantique permet d'alimenter des outils de science des données avec des données supplémentaires, rendant de ce fait les analyses plus précises. Pour donner un ordre

1. Un graphe est une structure de données offrant un pouvoir d'expression très général, constituée d'un ensemble de nœuds (ou sommets) reliés entre eux par des arêtes (ou des arcs).

de grandeur, on estime[2] à 80 % les données disponibles sous forme textuelle, non structurée, dans les organisations. D'un point de vue qualitatif, les données issues de l'analyse textuelles sont nouvelles, originales, car elles n'existent généralement pas ailleurs sous une forme déjà structurée. Elles permettant de mieux comprendre une situation, par exemple *pourquoi* les consommateurs sont satisfaits ou non des produits et services proposés, ou les raisons pour lesquelles des équipements industriels subissent des pannes.

On combine généralement trois tâches d'extraction d'informations (figure 1.1) : la **classification automatique**, la **reconnaissance d'entités** (ou de concepts) ainsi que la **détection de relations** entre ces entités. Ces tâches de haut niveau s'appuient elles-mêmes sur des tâches plus simples, comme reconnaître la langue d'un texte, découper ce texte en phrases et en mots, identifier la classe grammaticale d'un mot[3], passer de sa forme fléchie à sa forme de base[4]… Toutes ces tâches, et les difficultés rencontrées, sont détaillées dans le chapitre 4. Historiquement, elles étaient programmées par un développeur ou créées sous forme de règles par un linguiste ; les percées récentes en apprentissage automatique ont permis d'améliorer significativement la performance de la plupart de ces tâches, notamment avec l'utilisation de réseaux de neurones artificiels comme nous le verrons au chapitre 5.

**Figure 1.1** – Trois tâches d'extraction d'informations : classification, reconnaissance d'entités et extraction de relations.

Ces différentes tâches ne s'exécutent pas indépendamment les unes des autres. Elles peuvent collaborer pour se transmettre des éléments de contexte. Par exemple, la classification préalable d'un document, pour en calculer les thématiques principales, permettra aussi d'aider à reconnaître des entités en excluant des hypothèses : le mot « Hollande » sera interprété localement avec des sens différents si la thématique globale du texte concerne les Pays-Bas ou la présidence de la République française.

---

2. Cet ordre de grandeur est très largement cité par les analystes, les éditeurs de logiciels et les utilisateurs, qui cherchent tous à démontrer l'utilité de l'analyse de texte. Seth Grimes présente une analyse critique de cet indicateur dans [2]. L'apparition de l'Internet des objets (IoT) a créé des nouveaux canaux de données gigantesques, et contribue donc à diminuer ce pourcentage.

3. Est-ce qu'un mot est un nom, un verbe, un adjectif, un adverbe, un déterminant, une préposition… ? Cette tâche s'appelle l'étiquetage morphosyntaxique (*part-of-speech tagging*).

4. Par exemple, trouver l'infinitif d'un verbe conjugué ou le masculin singulier d'un nom au féminin pluriel (voir page 136).

De même, une entité reconnue d'une façon sûre, sans ambiguïté, pourra fournir un indice aidant à mieux classifier globalement un document.

### ◆ *Classification automatique de documents*

La **classification automatique** d'un document (ou d'une partie de ce document) revient à effectuer une compréhension *globale* de son contenu pour le ranger dans une ou plusieurs cases. En quelle langue est-il écrit ? De quel type de documents s'agit-il ? Faut-il le traiter en urgence ? Quelles sont ses principales thématiques ? Un mail est-il un spam ou non ?

Lors de l'assemblée générale des actionnaires tenue le 7 mai 2019 au Palais des Congrès, à Paris, Air Liquide a innové : les actionnaires présents dans la salle ont été invités à envoyer par SMS les questions qu'ils souhaitaient poser au Président. Les centaines de questions reçues ont été analysées en temps réel par un système de classification automatique qui les a catégorisées et regroupées en sujets pertinents. L'application a identifié les thématiques les plus importantes aux yeux des actionnaires en illustrant chacune d'elles avec les questions les plus représentatives. Cela représente un cas d'usage original de la classification automatique, au service de la démocratie actionnariale.

### ◆ *Reconnaissance d'entités*

La compréhension *locale* d'un texte consiste à « stabiloter » des mots ou groupes de mots pour reconnaître des concepts. En jargon de linguiste, cette tâche s'appelle la **reconnaissance d'entités nommées** (*named entity recognition* ou *NER*).

Le cas le plus simple revient à identifier dans un texte une information unitaire comme une date, un montant financier, un pourcentage, un numéro de téléphone, une adresse de mail, une URL, un numéro de plaque d'immatriculation, un numéro de sécurité sociale… On utilise généralement pour cela une expression régulière (*regular expression* ou *regex*), c'est-à-dire une description formelle de la suite de caractères cherchés : par exemple, un pourcentage sera décrit comme un nombre entier suivi du caractère pourcentage.

Un cas plus complexe consiste à reconnaître dans le texte des personnes, lieux, organisations… mais aussi – en fonction du prisme de l'analyse attendue – des molécules, des compétences RH, des pays ou tout autre type d'entité de granularité arbitrairement fine. Par exemple, plutôt que de reconnaître des personnes d'une façon générale, on peut souhaiter distinguer des politiques, des militaires, des diplomates, des sportifs, etc. Certaines entités existent en nombre fini : leur liste est connue à l'avance (par exemple les atomes, les planètes du système solaire, les pays).

Plusieurs difficultés linguistiques doivent être prises en compte : homonymie (*Orange* peut désigner une entreprise, une ville, mais aussi une couleur ou un fruit), polysémie (dans « Paris a décidé que… » Paris est une entité politique et non un lieu), synonymie (l'entreprise Orange est parfois aussi appelée « France Telecom » ou « l'opérateur historique » dans le contexte de l'industrie télécom en France).

Difficulté supplémentaire : pour éviter la répétition d'un nom, une entité n'est pas toujours explicitement nommée, mais désignée par un pronom ou une formulation alternative. Cette figure de style (appelée anaphore) nécessite de calculer à qui ou quoi correspond le pronom utilisé. Dans l'extrait d'article de presse *« Richard Ferrand propose la nomination d'Alain Juppé au Conseil constitutionnel. Le maire de Bordeaux et ancien Premier ministre succéderait au socialiste Lionel Jospin. Il a annoncé mercredi qu'il avait accepté cette proposition. »* les cinq éléments soulignés sont des anaphores qui font référence à la même personne politique[5].

### ◆ *Extraction de relations*

Des relations entre entités existent : une société nomme un dirigeant ; une société rachète une autre société ; un auteur publie un livre ; un pays nomme une personne à un poste d'ambassadeur… L'**extraction de relations** est plus complexe que celle des entités car une relation donnée peut être exprimée d'un grand nombre de façons. « X achète Y » peut aussi s'écrire « X rachète Y », « X fait l'acquisition de Y », « Y récemment rachetée par X », « Y nouvelle filiale de X » ou d'une trentaine d'autres façons. Des variantes peuvent décrire une situation proche, comme « X a pris une participation majoritaire dans Y ».

L'analyse d'un texte permet aussi de relier une entité à ses attributs ou propriétés. Par exemple, un pays a un nombre d'habitants, une superficie, un PIB, etc. Un ordinateur a des caractéristiques techniques comme sa capacité mémoire, son microprocesseur, etc. Notons que la datation de l'information est alors aussi un élément à prendre en compte, la valeur d'une propriété d'une entité variant dans le temps (comme la population d'un pays ou le poids d'une personne). Or l'interprétation d'une date apparaissant dans un texte est plus complexe qu'il n'y parait : s'agit-il de la date où le document a été écrit ou de la référence à une date passée ou future ?

### ◆ *Du sur-mesure ou du prêt-à-porter ?*

#### Analyseur générique

Comme souvent en traitement du langage, plusieurs approches sont envisageables pour effectuer une analyse sémantique : utiliser un **analyseur générique** disponible sur étagère, en créer un nouveau sur mesure bien adapté au corpus à traiter, ou choisir de combiner plusieurs analyseurs. Un point important à comprendre est qu'aucun analyseur ne donnera un résultat sans erreur : sur la plupart des tâches de haut niveau décrites ici, atteindre une qualité de 90 % est déjà très satisfaisant (nous verrons comment la mesurer page 264) et souvent équivalent au résultat d'un travail humain.

Si on souhaite traiter la plus grande variété de textes possibles, il faut privilégier un analyseur générique. La disponibilité croissante de corpus de très grande taille, annotés ou non (Web, news, romans, Wikipédia…), permet de créer des analyseurs génériques de qualité acceptable pour certaines tâches (notamment la classification

5. Le premier « il » de la dernière phrase pourrait faire référence aussi bien à Richard Ferrand qu'à Alain Juppé.

automatique) grâce à un apprentissage à large échelle : l'encyclopédie Wikipédia propose aujourd'hui des grandes quantités de texte dans la plupart des langues. Ces extracteurs génériques d'informations offrent l'intérêt de fonctionner sans aucun paramétrage, mais n'analysent correctement que des documents proches de ceux sur lesquels ils ont été entraînés. Si on utilise un analyseur de CV sur des articles de presse, il cherchera à reconnaître des compétences ou des talents RH sans même avoir conscience de ne pas être bien adapté aux documents traités, ce qui se traduira par une grande confusion dans le résultat produit.

#### Analyseur spécifique

Si on connaît à l'avance le corpus à traiter et qu'on privilégie la qualité de traitement, on doit définir un **analyseur spécifique** pour s'adapter au mieux aux spécificités du corpus, à sa phraséologie et aux types de concepts qui y apparaissent. Par exemple, pour l'analyse du Grand débat national lancé en 2019 par le gouvernement français, les thématiques abordées sont suffisamment spécifiques pour nécessiter la construction d'un analyseur sur mesure.

L'apprentissage automatique a fourni des nouveaux moyens pour faciliter une analyse sur-mesure, à un coût abordable, dans des délais maîtrisés. En début de projet, l'apprentissage profond permet de faire émerger les concepts pertinents à partir du corpus et d'amorcer rapidement une première organisation structurée des concepts (aussi appelé taxonomie), évitant ainsi le syndrome de la page blanche. Pendant la vie du projet, l'intelligence artificielle permet aussi de détecter les concepts nouveaux au fur et à mesure de leur apparition (par exemple des marques nouvelles ou des risques nouveaux) et d'enrichir la taxonomie existante avec un corpus qui évolue. Ces systèmes permettent aussi de visualiser des phénomènes émergents, en détectant des signaux faibles imperceptibles à l'œil nu, dès leurs prémices.

Enfin, une force de l'apprentissage profond est la capacité à traiter simultanément la plupart des langues. À partir du moment où un corpus de large taille est disponible pour une langue (Wikipédia ou historique de verbatim), un modèle de langage peut être appris automatiquement. Il devient alors possible de centraliser tous les verbatim au sein d'un même référentiel et ainsi de mesurer les phénomènes, quelle que soit la langue dans laquelle ils sont exprimés. C'est idéal, dans un monde globalisé, pour comprendre ce qui se passe dans une filiale étrangère ou adapter une offre à des spécificités culturelles.

### 1.2.2 Fouille de textes : extraire des connaissances d'un corpus

La **fouille de textes** (*text mining*) revient à effectuer l'analyse sémantique des documents d'un corpus pour alimenter un système de fouille de données (*data mining*). Ce dernier interprétera les résultats de l'analyse textuelle de façon à mettre en évidence des corrélations intéressantes, effectuer une analyse de séries temporelles, détecter des corrélations… La fouille de textes fonctionne mieux sur un corpus suffisamment homogène pour que le même analyseur puisse traiter tous les documents. Au pire, si la cohérence du corpus ne peut être garantie, l'extraction

d'information effectuée sera simplifiée pour se limiter à l'utilisation d'analyseurs génériques et à l'indexation classique d'un moteur de recherche (voir page 86). Le cumul de toutes les informations extraites des différents documents crée de nouvelles connaissances.

Le monde médical a rapidement vu l'intérêt des technologies de fouille de textes et de visualisation de données. Appliquées aux articles scientifiques, elles permettent d'afficher les termes biomédicaux en cooccurrence sous forme de réseaux graphiques. Cette cartographie donne un aperçu des relations possibles entre ces termes ; elle permet par exemple de voir des interactions entre médicaments et protéines qui resteraient sinon invisibles à l'œil nu, et de trouver les articles scientifiques pertinents.

Dans les applications de fouille de textes, on peut aussi citer les systèmes de veille utilisés par les services de renseignement, ou par les entreprises dans le cadre de leur activité d'intelligence économique.

### 1.2.3 Analyse d'opinions : comprendre les attentes d'un écosystème

Les humains expriment leur avis spontanément (par exemple en écrivant sur les réseaux sociaux ou en envoyant des mails de réclamation) ou en réponse à une sollicitation (notamment à la suite d'une enquête de satisfaction). Quelles opinions sont exprimées par celui ou celle qui s'exprime ? S'agit-il d'un sentiment positif, négatif, neutre ou mitigé ? Sur quoi porte précisément cet avis ? Quelles émotions se dégagent entre joie, colère, peur, surprise, tristesse, dégoût, confiance… ? Un consommateur qui écrit « *je suis surpris de ne toujours pas avoir reçu de réponse* » est-il d'ailleurs surpris, déçu ou en colère, ou ressent-il plusieurs de ces émotions simultanément ?

Le champ d'application de l'**analyse d'opinions et de sentiments** est très vaste. Il peut s'agir pour une marque d'être à l'écoute des consommateurs ; pour une direction des ressources humaines de mieux comprendre les attentes des collaborateurs et de mesurer la qualité de vie au travail ; pour un parti politique ou un gouvernement, d'inciter les citoyens à s'exprimer dans le cadre d'un débat ; pour une grande entreprise, de solliciter l'avis de ses actionnaires en direct pendant l'assemblée générale.

D'un point de vue technique, l'analyse d'opinions peut être vue comme un cas particulier d'extraction de relations. En effet, une opinion relie un locuteur qui s'exprime et l'objet du monde concerné (produit marchand, service proposé, action d'un politique…).

Ce type d'application créé des attentes élevées au niveau des directions marketing et communication. Dans le meilleur des cas, on arrive à classer correctement 90 % des verbatims, quand le contexte s'y prête, avec des documents homogènes, exprimant effectivement une opinion dans une langue que la machine maîtrise.

La tâche est en effet très subtile. Un verbatim comme « *je craque* » est positif ou négatif en fonction du contexte. De même, l'ordre des mots est essentiel pour faire le contraste entre « *rien ne vaut le Nutella* » et « *le Nutella ne vaut rien* ».

La gestion de la négation ne facilite pas toujours les choses. Elle est difficile à identifier pour un système d'analyse qui se limiterait au comptage des mots clés. Il faut tenir compte de doubles négations ou de leur absence. Quant à l'ironie, au sarcasme ou à l'hyperbole, on entre au-delà du champ de ce qu'une machine peut analyser avec succès de nos jours : la compréhension de ces phénomènes nécessite en effet d'avoir des connaissances du monde, extratextuelles.

### 1.2.4 Recommandation : faire le matching de contenus

Un **système de recommandation** (*matching*) utilise des technologies de type moteur de recherche pour mettre en correspondance une requête – généralement exprimée par quelques mots – et les documents les plus pertinents associés à cette requête. Il applique le même type de normalisation à la requête et au contenu des documents pour les représenter sous forme de vecteurs. Un calcul[6] permet alors de les comparer entre eux pour trouver les documents les plus appropriés ; ce fonctionnement, qualifié de modèle vectoriel, est détaillé page 86.

Classiquement, la limite de cette approche est d'utiliser les mots tels quels, sans prendre en compte les synonymes : une recherche ne renvoie que les documents contenant strictement les mots clés de la requête. Or, comme chaque requête ne contient que peu de termes, l'algorithme de similarité ne renvoie pas toujours des pages d'une grande pertinence. Des progrès technologiques récents (réseaux de neurones) permettent de lever ces contraintes.

Google a introduit en 2018 une fonctionnalité d'expansion sémantique appelée *neural matching* [4], fruit de cinq ans de R&D. La motivation de cette amélioration est de résoudre un problème bien connu : les mots que les gens utilisent pour effectuer une recherche sur internet sont souvent différents de ceux qui apparaissent dans les pages les plus pertinentes. Cette fonctionnalité améliorerait les résultats pour plus de 30 % des recherches, toutes langues confondues. En octobre 2019, Google a introduit le modèle BERT (voir page 216) dans la version américaine de son moteur de recherche ; la firme de Mountain View revendique une bien meilleure compréhension de certaines requêtes effectuées par les internautes, même en cas de requête totalement nouvelle. Rappelons que Google a la contrainte d'utiliser des solutions techniques généralistes, fonctionnant sur l'ensemble du Web, d'où son intérêt pour des modèles de langues universels.

#### ◆ *Aide au recrutement et à la recherche d'emploi*

Si on peut faire des hypothèses sur le corpus, il devient possible de mettre en place une expansion de synonymes spécifique qui donnera de biens meilleurs résultats. Par exemple, dans l'univers du recrutement, on cherche à trouver le meilleur CV correspondant à une offre d'emploi (quand on recrute) ou le meilleur poste pour un profil donné (quand on est en phase de recherche d'emploi).

---

6. Ce calcul de produit scalaire, parfois appelé *cosinus de Salton*, est le fondement du modèle vectoriel introduit par Gerald Salton [3].