

## **1. Compte rendu de lecture : « Une petite introduction au Traitement automatique des langues naturelles ».**

Le TALN (Traitement Automatique des Langues Naturelles) est l'« ensemble des recherches et développements qui visent à modéliser et reproduire, à l'aide de machines, la capacité humaine de » langage. Cela n'est pas aussi simple en réalité : en effet, le traitement automatique des langues naturelles posent de nombreux problèmes. Ici, on va plus spécialement s'intéresser aux langues naturelles sous leur forme écrite. L'automatisation des langues naturelles a deux raisons : pour pouvoir comprendre les « mécanismes de la communication humaine » en cherchant à la modéliser, et pour pouvoir traiter rapidement et efficacement des données qui concernent le langage. Le TAL est à la croisée de nombreux domaines (intelligence artificielle, linguistique, informatique...).

Faisons un bref historique : les premiers travaux du TAL ont porté sur la traduction automatique (à l'époque de la guerre froide). C'est en 1962 qu'a lieu la première conférence sur ce sujet, organisée par Bar Hillel. Depuis 1954, d'importants financements avaient donné lieu à de nombreuses recherches, qui n'auront pas beaucoup de réussite, et seront vite arrêtées car un groupe d'expert montre que la traduction automatique, à l'époque, coûte presque deux fois plus cher que la traduction humaine, et pour des résultats de moindre qualité.

Malgré cet échec, on a pourtant dans les années 1950 un développement de plusieurs idées fondamentales, avec les travaux de Harris et Chomsky notamment. C'est en 1956 que naît l'intelligence artificielle (IA), ainsi que, quelques années plus tard, des systèmes comme celui d'Eliza, qui mettent en œuvre des « mécanismes de traitement simple, à base de mots clés », même si cela s'applique alors à des domaines très réduits, avec des formes grammaticales prédéfinies, sans faire appel à des connaissances syntaxiques, sémantiques et encore pragmatiques.

Dans les années 70, on s'attache davantage au traitement de la sémantique, en essayant notamment de prendre en compte les connaissances générales communes aux interlocuteurs en plus de la connaissance d'un domaine précis ou de l'importance du contexte.

Le TAL se heurte à plusieurs difficultés, comme nous l'avons fait remarquer plus tôt, qu'on peut alors classer en deux catégories.

D'une part, les ambiguïtés de langage. Il y a en effet plusieurs interprétations possibles pour chacune des unités linguistiques, que ce soit au niveau des graphèmes (les mêmes lettres ne correspondent pas toujours aux mêmes sons), au niveau des terminaisons des mots (une même terminaison, le « s » par exemple, peut concerner à la fois un processus de conjugaison – la deuxième personne du singulier – et une inflexion – le pluriel des noms, des adjectifs...).

D'autre part, l'implicite. En effet, pour comprendre un énoncé, bien souvent le sens des mots ne suffit pas ; il faut aussi avoir une certaine connaissance du monde et de son fonctionnement qu'auront deux humains dans une conversation, mais que ne possèdera pas une machine. Cela est bien visible dans certains effets de style comme les ellipses ou les métaphores. Il faut alors, si l'on veut permettre aux machines de pouvoir comprendre ces énoncés, soit disposer d'une base de connaissance additionnelle, soit délimiter et limiter les textes analysés à des sujets ou domaines particuliers.

Il y a différents niveaux de traitement nécessaires pour effectuer l'analyse d'un énoncé et parvenir à sa compréhension.

Il faut d'abord séparer le texte en plus petites unités (les mots), puis leur apposer un sens lexical (traitement lexical), puis délimiter les mots qui se combinent entre eux dans un ensemble intermédiaire entre la phrase et le mot (le syntagme), ce qui correspond au traitement syntaxique. On essaye ensuite d'attribuer un sens à l'énoncé en entier, ce qui donne le traitement sémantique, puis arrive la dernière étape, le traitement pragmatique, qui replace l'énoncé dans le contexte où il a été produit.

En ce qui concerne la segmentation des unités, si elle peut paraître assez simple en français (chaque mot étant plus ou moins délimité par un espace), pour ce qui relève des séparateurs renvoyant à la ponctuation, ce n'est pas aussi simple. En effet, un même symbole peut renvoyer à plusieurs sens différents selon les langues (il faut alors prendre en compte tous les usages possibles de ces séparateurs).

En ce qui concerne les formats des documents électroniques possibles à traiter, on assiste à un développement des formats dits « enrichis », qui contiennent de nombreuses informations en plus du texte telles que la structure du document. Il y a encore peu de normes en vigueur sur ce domaine, ce qui rend leur exploitation plus difficile (il faudrait pouvoir traiter chacune de leurs spécificités).

Le traitement lexical consiste à « reconnaître dans chaque chaîne de caractère une ou plusieurs unités linguistiques, dotées de caractéristiques propres, telles que leur sens, leur prononciation, leurs propriétés morphologiques, syntaxiques).

L'accès lexical se fait soit par identification directe, soit par recherche de la forme dans un lexique pré-compilé. Cependant, une des propriétés principales du langage est la créativité : de nouveaux mots surgissent tous les jours, soit par emprunt à une autre langue, soit par création à partir de nouveaux mots existants.

Ceci nous amène donc à la morphologie : dans une unité telle que le mot, on peut décomposer une ou plusieurs unités plus petites que l'on appellera morphèmes. La morphologie est donc l'étude des morphèmes, et de deux processus bien distincts les impliquant : les processus flexionnels, qui sont imposés par les conditions syntaxiques d'utilisation du mot, et les processus dérivationnels, qui permettent la création de nouvelles formes à partir des formes déjà existantes.

A partir de ces processus, les professionnels du TAL ont pu mettre en œuvre de types de connaissances pour développer les analyseurs morphologiques : les règles de combinaisons morphématiques (c'est-à-dire les conditions sous lesquelles les associations de deux morphèmes sont permises ou non) et les règles d'ajustement orthographiques (qui permettent de préserver la forme orale d'un mot).

La création lexicale peut aussi être le résultat d'une composition, lorsque, à partir de mots composés, on obtient des mots ayant un sens particuliers qui n'est plus déductible de celui de leurs composants. Cela va alors augmenter sensiblement le nombre d'ambiguïtés possibles, puisque, dans un même groupe, on peut parfois avoir une interprétation figée et une interprétation littérale (notamment pour les mots composés qui sont simplement apposés côte à côte, et non reliés entre eux par une marque telle que le tiret).

La syntaxe est l'étude des contraintes qui agissent sur l'ordre ou les relations des mots entre eux et qui doivent être prises en compte lorsque l'on cherche à produire ou décrire des phrases grammaticales. Les contraintes sont là aussi de deux sortes : les propriétés

sélectionnées (telles que les règles d'accord) ou positionnelles (l'ordre des mots dans une phrase).

La grammaire correspond à la description de ces contraintes dans une langue donnée. Les connaissances syntaxiques peuvent être nécessaires dans le TAL pour pouvoir s'occuper d'une part de la correction automatique (les phrases peuvent n'être pas toujours grammaticales) et de l'autre pouvoir désambiguïser une phrase en 'étiquetant de manière syntaxique.

En effet, les énoncés ne sont pas de simples suites de mots, mais constitués de groupes de mots intermédiaires, les syntagmes. Le but d'une analyse syntaxique est d'associer à chaque énoncé sa structure en constituants. Plusieurs éléments prouvent l'existence de cette structure syntaxique : les variations paradigmatiques possibles entre différents constituants pour un même énoncé, les contraintes qui portent sur les déplacements de constituants dans certaines structures (lorsque l'on passe de la voix active à la voix passive par exemple) et enfin le test de la conjonction, ce qui n'est pas applicable à n'importe quel groupe de mots. Ainsi, l'analyse syntaxique revient à associer à une structure linéaire (la phrase) une structure hiérarchique rendant compte des relations entre unités la constituant.

Traditionnellement, le résultat de l'analyse syntaxique est représenté sous forme d'arbre. Cependant, l'analyse syntaxique est elle aussi confrontée à quelques problèmes : l'ambiguïté lexicale tout d'abord (une même forme peut appartenir à différentes catégories syntaxiques). Cette ambiguïté peut être dans la langue ou ancrée au contexte, comme on peut le voir avec les nombreux phénomènes de translation (qui rendent ambigus les adjectifs par exemple en les employant comme noms communs...). Des verbes peuvent aussi correspondre à des sens différents selon leur sous-catégorisation, ce qui peut aussi donner lieu à des phrases ambiguës.

Passons maintenant au traitement sémantique : cette dernière se préoccupe du sens des énoncés. Dans certains cas, le sens peut dépendre étroitement du contexte (notamment avec les déictiques). On construit généralement le sens d'une phrase d'après le sens trouvé des différents constituants mis bout à bout.

La représentation conceptuelle est la représentation d'un mot qui est stockée dans le lexique. La représentation fonctionnelle est la fonction sémantique d'un mot, cependant cette dernière prend peu en compte les notions de nuance ou de temporalité.

L'analyse syntaxique, et surtout la détermination de la structure syntaxique obtenue est nécessaire pour accéder au sens. Cependant, parfois, l'aide du contexte semble nécessaire. On arrive alors au niveau de la pragmatique. Cette dernière porte sur les opérations logiques déclenchées par les attitudes des locuteurs vis-à-vis des énoncés.

Ces opérations peuvent être du type de l'inférence logique (à partir du sens d'un énoncé, on effectue une déduction en utilisant une connaissance contextuelle). Il faut alors que tout énoncé soit pertinent pour permettre le traitement pragmatique. On peut également avoir des ambiguïtés pragmatiques, lorsque le sens d'un énoncé est parfaitement clair, mais que l'attitude que le locuteur veut dégager à travers cet énoncé ne l'est pas.

Pour représenter la pragmatique, on va faire appel à tout ce qui relève de la logique et de la planification, notamment en utilisant des « dialogues finalisés », c'est-à-dire orientés vers un but, en tenant compte à chaque instant des connaissances et intentions de l'interlocuteur.

Concernant le TAL, la pragmatique répond à deux besoins : celui de rendre les interfaces plus ergonomiques, et celui de pouvoir traiter des données de manière de plus en plus intelligente. Ces applications peuvent alors être classées en trois grandes familles.

La première est celle du traitement documentaire : elles cherchent à faciliter le traitement par les hommes des ressources disponibles en langage naturel de plus en plus nombreuses (comme la traduction automatique, la recherche de documents dans une base documentaire, le classement ou l'indexation de documents, ce qu'on appelle routage, la lecture automatisée de documents et enfin l'analyse de documents relatifs à un corpus donné.). Les thésaurus, et les reconnaissances de mots composés, permettent d'améliorer les outils d'indexation ou de recherche documentaire.

La deuxième est la production de documents, qui se manifeste avec le développement des claviers « auto-correcteurs », la reconnaissance optique de caractère, les correcteurs stylistiques, l'apprentissage assisté de langues par ordinateur et la génération automatique de documents à partir de spécifications formelles.

Enfin, la troisième et dernière les interfaces naturelles, permet l'interrogation de bases de données ou de moteurs de recherche sur internet, et toutes sortes d'interfaces vocales.

En conclusion, on peut dire que le domaine du TAL est encore un domaine émergent ou de nombreux progrès restent et peuvent être accomplis, même si de nombreuses avancées ont déjà permis de résoudre de nombreux problèmes et d'obtenir certains outils relativement performants. Les connaissances linguistiques, parfois encore trop peu formalisées pour être utilisées, expliquent l'exploitation de plus en plus fréquente de corpus pour obtenir ces informations manquantes, et l'utilisation de systèmes d'apprentissage automatique, encore en développement.