INITIATION AU TAL

Devoir 3

Chapitre 1 Introduction

En 1950, Turing, le père fondateur de l'informatique, prédit que « dans 50 ans », les ordinateurs auront acquis la capacité de parler.

Aujourd'hui, les ordinateurs battent les grands maîtres d'échecs à plate couture, mais ils n'ont toujours pas les compétences langagières d'un enfant de 5 ans.

Une des premières conférences réunissant les pionniers de cette discipline, dans les années 50, portait sur la « traduction automatique », mais l'histoire de ces tentatives est restée longtemps une succession d'échecs et de déconvenues. C'est l'histoire, en fait, d'une prise de conscience pour les informaticiens de ce que les linguistes savaient depuis longtemps : l'extrême complexité des langues naturelles.

Dans les années 70 et 80, les « systèmes experts » sont à l'honneur. Le thème général du « traitement automatique du langage naturel » (TALN), lui, franchit plus rarement les frontières des laboratoires. A l'image de ce qui se fait avec les « systèmes experts », on essaie alors de ramener l'usage du langage à des règles symboliques.

Les années 90 voient arriver des changements considérables : banalisation des ordinateurs personnels, avec des capacités de stockage et de traitement en progression exponentielle, apparition du Web. L'« ingénierie linguistique ».

Chapitre 2 Traitement Automatique du Langage Naturel

1 Introduction

On dénombre actuellement environ 5 000 langues différentes, dont beaucoup sont en voie de disparition faute de locuteurs. Tout être humain normalement constitué et inséré depuis sa naissance dans un groupe social est capable, vers l'âge de 5 ans, de tenir une conversation courante dans sa langue maternelle. Aucun singe et aucun ordinateur ne peut en faire autant. Parler est « le propre de l'homme ».

Pour désigner les langues humaines, on parle « langues naturelles ». Les langues naturelles s'opposent aux « langues artificielles » ou « formelles ». On peut classer parmi les « langues naturelles » les créations linguistiques intentionnelles comme l'Espéranto ou le Volapük, ou encore certaines « langues des signes » inventées spécialement pour répondre à un besoin : la communication.

L'objet de la linguistique est l'étude scientifique des langues naturelles et, à travers elles, du langage en tant que « faculté de langue » dans l'espèce humaine. Les linguistes ne sont pas nécessairement polyglottes ; ils cherchent à comprendre les principes qui régissent les langues. La linguistique n'est pas prescriptive.

Pour cette étude, l'informatique joue un rôle via le domaine du traitement automatique du langage naturel. C'est une discipline scientifique récente, qu'il ne faut pourtant pas réduire à la simple utilisation d'ordinateurs et de programmes. On peut dire que les fondements de l'informatique sont doubles :

- le codage des données à l'aide d'éléments discrets (les fameux 0/1)
- le codage effectif des traitements à l'aide d'algorithmes.

C'est par ce biais qu'on va aborder le traitement automatique du langage.

Maîtriser une langue requiert la manipulation de nombreuses données, et la mise en œuvre de nombreux traitements. Les linguistes les ont progressivement mis à jour. Le TALN est né de leur interaction.

2 Histoires croisées de la linguistique et de l'informatique

Les dates suivantes concernent soit l'une de ces disciplines exclusivement : (l) pour « linguistique » et (i) pour « informatique » soit les deux (ce que l'on marquera par (il) pour « informatique linguistique ».

- Avant le XVIII^{ème} siècle : les précurseurs de la linguistique sont les auteurs de grammaires descriptives d'une langue donnée (l), les précurseurs de l'informatique sont les mathématiciens qui décrivent des méthodes générales de calcul et les inventeurs de « machines à calculer » mécaniques comme Pascal et Leibniz (i).
- 1660 : publication de la « Grammaire générale et raisonnée » d'Arnaud et Lancelot.
- Aux XVIII^{ème} et XIX^{ème} siècles : c'est la reine de la linguistique comparative et historique.
- 1916 : (l) publication posthume. Saussure introduit plusieurs distinctions et concepts importants :
  - Il caractérise le langage comme la construction sociale d'un système de signes.
  - Il considère que le langage, en tant que faculté générale de s'exprimer au moyen de signes, se distingue de la parole.
  - Il distingue les dimensions diachroniques du langage.
  - Il distingue deux axes d'analyse d'un discours, en tant que suite de signes : l'axe syntagmatique et l'axe associatif ou paradigmatique
- Alan Turing (1912-1954): mathematician angles. En 1936, il propose le dispositif appelé « machine de Turing ».
- 1945 : Von Neumann (1903-1957), mathématicien et physicien américain d'initie dans un rapport le plan de construction des ordinateurs, tels qu'ils sont encore conçus de nos jours
- 1952 : première conférence sur la traduction automatique, organisée au MIT (Massachusetts Institute of Technology) par Yehoshua Bar-Hillel (1915-1975).
- André Martinet (1908-1999) : (l) linguiste français, connu notamment pour avoir caractérisé les langues naturelles par la propriété de la double articulation.

Toutes les langues humaines sont doublement articulées parce qu'elles combinent des éléments (discrets) à deux niveaux différents :

– la « première articulation » est celle qui permet la combinaison « d'unités douées chacune d'une forme vocale et d'un sens » qu'il appelle des « monèmes ». On peut dire que ce niveau est celui de la syntaxe.

– la « deuxième articulation »

- 1966 : Deux programmes célèbres datent de cette époque : « Eliza » et « Student » de Bobro.
- Noam Chomsky linguiste et activiste politique américain, selon lui, les hommes disposent à la naissance d'un « organe du langage » de nature mentale. Il a élaboré de nouveaux concepts et promu plusieurs distinctions inédites :
  - Il distingue la compétence linguistique de la performance. La linguistique étudie le langage indépendamment de son substrat biologique
  - Il considère que le but de toute théorie linguistique est l'explication des jugements de grammaticalité.

- 1972 : l'informaticien Terry Winograd présente son programme intitulé SHRDLU permettant des interactions langagières avec un ordinateur sur un domaine restreint à un monde de blocs.

3 Les niveaux d'analyse du langage

La chronologie précédente a montré que ce n'est que progressivement que les Sciences du langage ont précisé leur objet d'étude. Dans le schéma, les principales unités d'analyse des plus simples aux plus complexes. Chacun des « nœuds » du schéma est constitué d'un ensemble de données discrètes.

Ce schéma permet de mettre en évidence plusieurs des spécificités des langues naturelles évoquées précédemment :

- la dimension « verticale » du schéma est l'axe syntagmatique, tandis que sa dimension « horizontale » correspond à l'axe paradigmatique.
- la « double articulation » du langage se retrouve dans les deux niveaux de traitements combinatoires successifs qui occupent l'axe syntagmatique.
- ce schéma opère une claire distinction entre deux niveaux souvent confondus : celui de la sémantique lexicale, qui étudie le sens d'unités individuelles, et la sémantique propositionnelle qui étudie le sens d'énoncés complets.

Chacun des deux niveaux de l'axe syntagmatique traduit une combinatoire ouverte d'éléments discrets. Ce dispositif original diffère fondamentalement de codages de type analogique. Comme le remarque le psycholinguiste canadien Stephen Pinker, les deux systèmes les plus sophistiqués de transmission d'information qui ont été sélectionnés par la nature, reposent tous les deux sur des unités discrètes.

Le niveau de l'association entre « unités significantes » et « sémantique lexicale » est un jalon important de notre schéma. Il constitue une étape fondamentale de l'acquisition de leur langue maternelle par les enfants, aux alentours de leur première année de vie.

Pourtant, le critère de la « double articulation » est probablement insuffisant pour distinguer à lui tout seul les langues humaines de langages d'un autre genre apparus récemment. Pour donner aux ordinateurs l'incroyable pouvoir expressif de ces langues, à savoir leur capacité à référer à ce qui leur est extérieur, à dire des choses sur le monde, il reste à réduire cette description du monde à un calcul.

4 La chaîne de traitements “standard”

Pour comprendre comment les humains utilisent une langue naturelle, il faut savoir comment ces connaissances sont exploitées. Cela ne relève plus de la linguistique, mais de la psychologie, voire de la psycholinguistique. La figure 2.2 propose une chaîne de traitements qui a une certaine plausibilité.

Dans ce schéma les données figurent dans des ovals, tandis que les traitements sont représentés dans des rectangles. Les deux principaux rectangles, intitulés respectivement « analyses » et « synthèses », correspondent aux deux tâches principales accomplies par les locuteurs d'une langue. Le fait de bien les séparer provient de l'observation de patients souffrant de lésions cérébrales qui affectent en particulier une de ces compétences et pas l'autre.

Suivant ce schéma, comprendre un énoncé revient donc à le transformer, via une « analyse », en une représentation interne, tandis que pour en produire ou en générer un, il faut traduire linguistiquement une telle représentation via une « synthèse ».

Chacune de ces tâches nécessite de prendre en compte l'ensemble des niveaux d'analyse identifiés précédemment, mais dans un ordre différent et en faisant chaque fois des hypothèses différentes sur ce qui est connu et ce qui doit être accompli.

Pourtant, le schéma de la figure 2.2 n'est instancié dans sa totalité dans aucun système informatique existant. C'est un cadre idéal théorique, très influencé par une conception cognitiviste de l'esprit humain, où les « représentations internes » sont en général de nature symbolique.

Après avoir longtemps tenté d'imiter le fonctionnement de l'esprit humain, les chercheurs du domaine essaient plutôt désormais d'exploiter au mieux les capacités de mémoire et de calculs de leurs machines.