Analyse de l’écriture

# L’ontologie

En informatique, une ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné. Plus simplement, on peut aussi dire que l' « ontologie est aux données ce que la grammaire est au langage ».

Les concepts sont organisés dans un graphe dont les relations peuvent être :

* des relations sémantiques ;
* des relations de subsomption.

L'objectif premier d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné, qui peut être réel ou imaginaire.

Les ontologies sont employées dans l'intelligence artificielle, le Web sémantique, le génie logiciel, l'informatique biomédicale ou encore l'architecture de l'information comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde. Les ontologies décrivent généralement :

* individus : les objets de base ;
* classes : ensembles, collections, ou types d'objets1 ;
* attributs : propriétés, fonctionnalités, caractéristiques ou paramètres que les objets peuvent posséder et partager ;
* relations : les liens que les objets peuvent avoir entre eux ;
* événements : changements subis par des attributs ou des relations.

La sémantique définie par signifier, indiqué où marque, étudie-les signifiés, soit ce dont parle un énoncé données. Grace au père fondateur de la sémantique moderne, Michel Jules Alfred Bréal, le domaine linguistique reconnait une évolution prépondérante de nos jours.

En particulier, la sémantique possède plusieurs objets d'étude :

* la signification des mots composés ;
* les rapports de sens entre les mots (relations d'homonymie, de synonymie, d'antonymie, de polysémie, d'hyperonymie, d'hyponymie, etc.) ;
* la distribution des actants au sein d'un énoncé ;
* les conditions de vérité d'un énoncé ;
* l'analyse critique du discours ;
* la pragmatique, en tant qu'elle est considérée comme une branche de la sémantique.

L’analyse sémantique d’un énoncé se concentre sur la caractérisation de ses éléments de bases, les mots, leurs propres constituants. Il s’intéresse à la structure en observant le mécanisme propre à la construction du sens, à voir mot pour mot.

Annotation sémantique

# Introduction

Un des principaux enjeux actuels du Traitement Automatique du Langage Naturel (TALN) est de capter l’information portée par un texte et d’accéder à son sens. Ceci passe, entre autres, par le traitement des unités linguistiques à forte valeur informative (et/ou référentielle). Le cours de l’histoire, ou plutôt de la recherche, a voulu que l’on désigne un certain nombre d’entre elles sous le nom d’ « entités nommées » (EN). Ces dernières correspondent traditionnellement à l’ensemble des noms propres présents dans un texte, qu’il s’agisse de noms de personnes, de lieux ou d’organisations, ensemble auquel sont souvent ajoutées d’autres expressions comme les dates, les unités monétaires, les pourcentages et autres.

La reconnaissance et la catégorisation de ces entités apparaissent ainsi comme une tâche fondamentale pour diverses applications de TALN participant de l’analyse de contenu, à l’instar de la recherche et l’extraction d’information, le question-answering ou encore l’intelligence économique. Considérée à tort comme triviale, elle a fait cette dernière décennie l’objet d’une attention plus soutenue, consécutivement à la récente multiplication des demandes issues du secteur industriel, et à la définition de tâches spécifiques lors des dernières compétitions MUC (Message Understanding Conference) et ACE (Automatic Content Extraction).

Les recherches menées jusqu’à ce jour ont ainsi permis l’élaboration de systèmes de reconnaissance d’EN relativement performants, permettant d’identifier dans des textes de nature journalistique des noms relevant de types généraux tels que « personne », « lieu » et « organisation », et ce avec un taux de F-mesure2 dépassant généralement les 0,90.

# Reconnaissance d’entité nommés

Compte tenu de la variété et de l’étendue (linguistiquement parlant) des unités pouvant faire partie de cet ensemble, un exercice de définition des EN n’est bien sûr pas des plus faciles. Considérer d’un point de vue « linguistique théorique » des critères et mécanismes distinctifs des EN et, partant, s’interroger sur leur statut référentiel constituent ainsi une perspective de recherche significative autour de la problématique des EN.

La seconde orientation est quant à elle plus « pratique », touchant aux objectifs et méthodes de traitement automatique des EN. L’approche proposée repose sur, d’une part, la construction dynamique d’une ressource lexico-sémantique dédiée aux EN, proposant pour ces dernières des étiquettes sémantiques fines et, d’autre part, la combinaison de cette ressource (permettant une annotation fine) avec un système standard de reconnaissance d’EN (proposant des étiquettes générales classiques) afin d’obtenir une annotation enrichie, ou double annotation, des EN.

L’originalité de cette approche réside dans l’exploitation de relations syntaxiques profondes lors de la construction de la ressource d’EN, dans l’annotation à l’aide de groupes d’étiquettes plutôt qu’à l’aide d’étiquettes et dans la mise en œuvre d’une double annotation des EN offrant des informations de niveau tant général que particulier

Les phénomènes de pluralité interprétative sont légion en langue naturelle, concernant de nombreux types d’unités linguistiques (noms, verbes, etc.) et jouant à tous les niveaux (morphosyntaxique, lexical, phrastique). Largement décrits et étudiés pour les unités lexicales classiques, les changements, transferts ou superpositions de sens le sont en revanche très peu pour les unités de type EN. Or celles-ci semblent au contraire être régies par les mêmes phénomènes. Considérons les énoncés suivants :

(1) Orange a invité M. Dupont.

(2) Leclerc a fermé ses magasins en Rhône-Alpes.

(3) La France a signé le traité de Kyoto.

l’homonymie (1) à la polysémie (2), en passant par la métonymie (3), force est donc de constater que les EN n’échappent pas aux phénomènes d’ambiguïté et sont, à l’instar des autres unités lexicales, polyréférentielles. Parallèlement à ces exemples « linguistiques », la réalité du traitement automatique de l’information révèle peu ou prou la même chose. Une interrogation du moteur de recherche Google pour les entités Orange et Leclerc propose (sur les dix premières réponses) pour l’une, des renvois aux référents « opérateur » et « ville » et, pour l’autre, des renvois aux référents « centres d’achat », « général », « char », et « homme d’affaires ». Cette réalité des choses conduit à penser l’annotation des entités nommées en des termes différents et à mettre en œuvre de nouveaux traitements.

## Approche

L’objectif est d’associer une information sémantique fine aux EN (construction de la ressource) tout en conservant le fruit des travaux réalisés jusqu’à aujourd’hui (combinaison avec un système classique) afin d’offrir une double annotation des entités nommées et de pouvoir désambiguïser certaines de ces unités.

La première étape correspond à la construction de la ressource d’EN. Cette ressource permet d’associer, pour chaque entité, une ou plusieurs étiquette(s) sémantique(s) fine(s) rendant compte de certaines caractéristiques du ou des référents possibles de l’entité. En d’autres termes, elle permet, idéalement, d’associer à l’entité Leclerc les étiquettes « société, homme d’affaires, char, général ».

Par son mode de construction, cette ressource pallie certains écueils de la recherche actuelle sur les EN tout en prenant en compte les particularités de ces unités. En effet, constituée automatiquement à partir de corpus, cette ressource peut tout d’abord être construite à moindre frais (or chacun connaît le coût de construction de lexiques spécialisés ou le coût de l’annotation manuelle d’un corpus d’apprentissage).

## Méthode

Ce que nous appelons ressource d’EN est une liste d’EN avec pour chacune d’elles une liste d’étiquettes sémantiques fines potentielles (par exemple les étiquettes « porte avions », « maréchal », « avenue », « hôpital » pour l’entité nommée Foch) provenant d’un corpus.

Le principe général de construction de cette ressource est l’identification dans le corpus de mots ou groupes de mots étant en relation avec les EN et pouvant servir d’étiquettes sémantiques. Afin de repérer et d’associer pertinemment entités et étiquettes.

Le processus de construction de la ressource se déroule en trois étapes :

* identification des relations syntaxiques pertinentes permettant d’associer des entités avec des étiquettes,
* construction « effective » de la ressource
* gestion des étiquettes par le calcul de cliques. Il convient de détailler chacune de ces étapes ; nous précisons auparavant les données et les outils utilisés.