La reconnaissance des entités nommées par les moteurs de recherche

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

De nombreux moteurs de recherche majeurs basent aujourd'hui une partie de leurs algorithmes sur la détection des entités nommés : nom de personne, d'entreprise, de lieu, etc. Mais peut-on facilement définir ce qu'est une entité nommée et comment les moteurs les reconnaissent-ils dans les documents qu'ils indexent ? Où en sont les chercheurs dans ce domaine ? Les travaux actuels sont-ils fiables ? Qu'en est-il dans le cadre d'une approche multilingue ? Cet article aborde tous ces sujets et tente de faire le point sur un pan important du Web sémantique, exploré aujourd'hui par Google et ses concurrents...

Pour construire un moteur de recherche capable de renvoyer des résultats pertinents, savoir reconnaître qu'un terme représente un nom de personne, une raison sociale d'entreprise ou un nom de lieu représente un atout certain. Ce problème est pris en charge par les techniques de "reconnaissance d'entités nommées" ("*Named entities recognition*" (NER) en anglais). Certains spécialistes en extraction de l'information ont annoncé voila plusieurs années que leurs méthodes étaient à présent "mûres", annonçant savoir reconnaître plus de 85%, voire plus de 90% des "entités nommées" dans un texte.

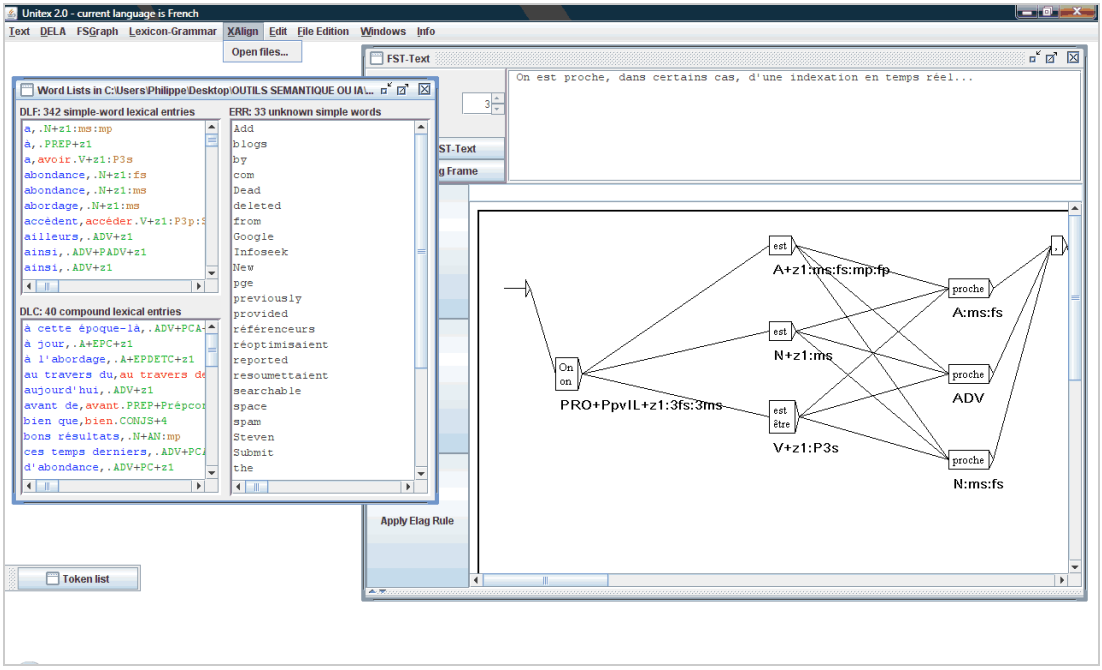
En réalité, nous verrons que certains problèmes sont loin d'être résolus, même si les progrès faits dans ce domaine sont rapides, spectaculaires, et leurs applications multiples. Et les principales avancées dans ce domaine pour les moteurs de recherche sont accélérées par la compétition entre les chercheurs de Yahoo, Microsoft et Google. Mais nous verrons que ce sont les équipes de Google qui se sont avérées depuis quelques mois extrêmement prolifiques en brevets et en publications scientifiques, et nous commençons à voir apparaître dans le fonctionnement de Google de nombreuses applications de leurs recherches sur les entités nommées.

Qu'appelle-t'on exactement "entité nommée" ?

Le terme a été défini pour la première fois lors de la sixième conférence MUC (*Message Understanding Conference*) en 1996. La recherche sur la reconnaissance des entités nommées à d'abord été stimulée par la compétition entre équipes de recherche en vue des conférences MUC, avant que des perspectives d'applications concrètes ne vienne prendre le relais. D'autres compétitions entre chercheurs sont organisées dans le cadre des conférences ACE (*Automatic Content Extraction*), ou des campagnes DUC en Europe. Mais il n'existe pas de définition universellement reconnue des entités nommées.

En règle générale, on désigne sous le vocable "entité nommée" une unité lexicale (un mot, ou un groupe de mots) qui fait référence à une entité (une chose, une personne, un lieu, une date, une mesure...) du monde concret. Cette unité lexicale constitue le nom de cette "entité", qui devient ainsi une entité nommée. Dans la pratique, une entité nommée sera le plus souvent unique, mais peut avoir plusieurs noms (London/Londres, Paris/La capitale de la France/La ville lumière).

On pourrait se dire qu'il s'agit d'un nom savant pour désigner les "noms propres". En fait les deux notions sont bien différentes. Si dans la plupart des cas, les notions de "noms propres" et "d'entités nommées" se recouvrent, il existe quand même des différences. Par exemple, l'expression "Le président de la République Française" est bien une entité nommée, car l'expression désigne aujourd'hui clairement une personne physique. Mais ce n'est pas un nom propre...



L'outil UNITEX en action : un logiciel Open Source utilisé par de nombreux chercheurs en TALN pour détecter des expressions candidates au statut d'entités nommées

Identifier une entité nommée n'est pas une tâche triviale

Pour identifier une entité nommée, on peut s'appuyer sur différents critères, mais certaines unités lexicales ne répondent à aucun de ces critères, et sont des entités nommées. D'autres répondent à la plupart de ces critères, et n'en sont pas !

On peut s'appuyer sur la **graphie** : dans la plupart des langues, les noms propres s'écrivent avec une première lettre en majuscule. Mais des exceptions existent, toutes les entités nommées ne sont pas des noms propres, et ne suivent pas cette règle ("le pape" et non "le Pape"). La règle n'est pas universelle (en allemand, tous les noms commencent par une majuscule, qu'ils soient des noms communs ou des noms propres) et la règle n'est jamais régulière ("Président de la République" ou "président de la république" ?) notamment en ce qui concerne les entités composés de plusieurs mots.

On peut s'appuyer sur les **cinq Q** (ou les cinq W en anglais) : les entités nommées sont en principe des réponses à des questions commençant par "*qui, quoi, quand, comment, où*". Sauf que beaucoup de noms communs répondent aussi à ce genre de critères.

On peut aussi remarquer que les entités nommées sont souvent **invariantes** quand on change de langue : "Mick Jagger" s'appelle "Mick Jagger" dans toutes les langues. Mais ce critère souffre de nombreuses exceptions : Paris/Parigi, London / Londres etc...

Autre critère possible : les entités nommées désignent en général une **chose unique** (une seule personne, un seul lieu, une seule entreprise). Mais si "Jean Martin" désigne dans un contexte bien déterminé un seul individu, il existe plusieurs personnes qui s'appellent "Jean Martin". Et l'unité lexicale "iPhone 3G" désigne un grand nombre d'objets. Une entité nommée peut donc avoir aussi bien une portée générique, que spécifique.

Une entité nommée peut aussi répondre au à la définition d'un "désignateur rigide" (selon la théorie de Kripke). Un "désignateur rigide" désigne un seul et unique objet bien déterminé dans tous les mondes possibles. George W. Bush est un désignateur rigide pour un ancien président des Etats Unis. Mais "l'actuel président des Etats Unis" est une entité nommée qui désigne une seule personne (Barack Obama) et ce n'est pas un "désignateur rigide" selon la définition de Kripke (Voir l'ouvrage *Naming and Necessity* qui regroupe trois conférences du Philosophe Saul Kripke à Princeton en 1970).

Bref, non seulement la définition est "floue", sans bases théoriques (le concept d'entités nommées est né de tentatives de résolution de problèmes pratiques en traitement automatisé du langage, pas de la théorie linguistique dont le caractère incomplet est ainsi violemment mis en lumière), mais il est de plus difficile de définir des critères identifiant clairement ce qui est, ou n'est pas, une entité nommée.

La reconnaissance des entités nommées bute sur les problèmes classiques en linguistique

Le problème de la reconnaissance des entités nommées connait par ailleurs d'autres difficultés que l'on rencontre dans toute tâche de *traitement automatisé du langage naturel* (TALN) et dans les moteurs de recherche. *TALN* est un acronyme. La reconnaissance des sigles, des acronymes, et des abréviations est également un problème classique de reconnaissance d'entités nommées, car les acronymes désignent en principe des entités nommées.

Tout d'abord, dans un texte peut figurer une graphie erronée de l'entité nommée : "Jean Dupon**d**" au lieu de "Jean Dupon**t**".

Ensuite, il peut y avoir des problèmes d'homonymie (il y'a plusieurs Jean Dupond à Paris), et de métonymie : dans l'expression "Paris a froid, Paris a faim", Paris ne désigne pas la capitale, mais les habitants de la capitale. Et dans l'expression "monter à la capitale", le nom commun "capitale" est en fait une entité nommée qui désigne... Paris.

Par ailleurs, la plupart des entités nommées sont en fait polysémiques : elles ont plusieurs sens ! "Charles de Gaulle" désigne un porte avions, un aéroport, une place de Paris, de nombreuses rues et avenues et bien sûr le général ! Seul le contexte peut permettre de lever les ambiguïtés sur le "Charles de Gaulle" dont on parle.

Et enfin, la longueur du groupe de mots à prendre en compte est un problème difficile, qui génère beaucoup de faux positifs si l'on n'y prend pas garde : l'abbé Pierre Dupont est-il un abbé qui se prénomme Pierre, ou l'abbé Pierre ? Une analyse syntaxique permet parfois de résoudre ce genre de problèmes, doublée de la prise en compte du plus long syntagme comme l'entité nommée la plus probable.

Ensuite, il existe des cas dans lequel le contexte ne suffit pas à lever les ambiguités : "Hilton, Paris" fait il allusion à la bimbo, ou à l'hôtel Hilton de Paris ?


[company]Google was co-founded by [person]Larry Page and [person]Sergey Brin while they were [vocation]students at [school]Stanford University and the company was first incorporated as a privately held company on [month]September 4, 1998. The initial public offering took place on [month]August 19, 2004, raising US$1.67 billion, implying a value for the entire corporation of US$23 billion. [company]Google has continued its growth through a series of new product developments, acquisitions, and partnerships. [association]Environmentalism, philanthropy and positive employee relations have been important tenets during the growth of [company]Google. The company was being identified multiple times as [magazine]Fortune [city]Magazine's #1 Best Place to Work[4] and as the most powerful brand in the world, according to [person]Millward Brown[5]. The Company describes its mission as follows: " [company]Google's mission is to organize the world's information and make it universally accessible and useful. "

—[company]Google, [job_title]Corporate Information[6]

The unofficial company slogan, coined by former employee and Gmail's first engineer[7] [person]Paul Buchheit, is "Don't be evil"[8][9][10]. Criticism of [company]Google includes concerns regarding the privacy of personal information, copyright, censorship and discontinuation of services[11].

Un texte annoté à l'aide de l'outil Yooname : l'outil reconnait un grand nombre d'entités et sait les catégoriser. Il est intéressant d'identifier les entités qui sont mal catégorisées, et pourquoi il se trompe (exemple : Corporate Information ->Job Title)

Et catégoriser les entités nommées est moins simple qu'il n'y parait

Au fil des compétitions MUC, de nombreuses conventions ont été adoptées pour classer les entités nommées dans des catégories bien précises. C'est ainsi que sont apparues les conventions ENAMEX (noms de personnes, noms de villes), TIMEX (date, heure) et NUMEX (montants financiers, pourcentages). Mais en réalité, de nombreuses entités nommées ne rentrent pas dans ces cases, et il existe de nombreuses cas "hors classification" ou que l'on est obligé de placer dans plusieurs catégories à la fois.

Sur un plan pratique, et en particulier lorsque l'on veut utiliser la reconnaissance des entités nommées dans un moteur de recherche, il devient indispensable d'utiliser des catégorisations beaucoup plus fines, ce qui a tendance à rendre le problème de la catégorisation encore plus délicat.

Comment reconnait-on les entités nommées ?

On utilise en général des combinaisons de plusieurs approches :

- **l'identification de termes "inconnus"** : les textes sont tokenizés, confrontés à des dictionnaires de morphèmes pour identifier les candidats au statut d'entités nommées. Certaines entités nommées sont détectées directement à l'aide de "dictionnaires" de noms propres, de noms de lieux etc. Un morphème représente toute forme que peut prendre un verbe conjugué, ou un mot en fonction de son genre, de son nombre ou de son cas pour les langues à déclinaisons.

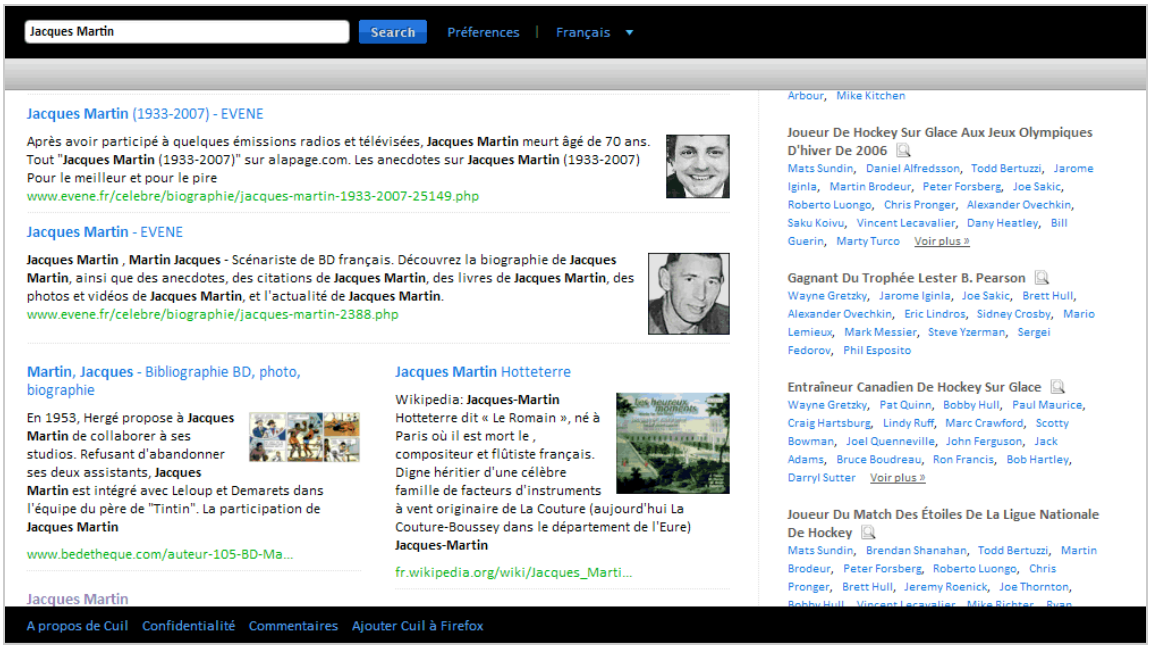
- **la reconnaissance de séquences caractéristiques** : un nom précédé de *Mme* et commençant par une majuscule est probablement un patronyme, un nom précédé de *SARL* une raison sociale. Le web fourmille par ailleurs d'exemples de collections de page qui ont toujours la même structure : si la balise <h1> contient "*Les Misérables (Victor Hugo)*" et que l'on est capable de reconnaître un titre de livre et son auteur grâce à une base de données, alors il sera facile à grands coups d'expressions régulières de reconnaître "*Les Fourmis (Bernard Werber)*", même si Bernard Werber n'est pas dans la base.

- **une identification semantico-linguistique** : une analyse syntaxique aidée de bases de données d'identifiants caractéristiques permet d'identifier les expressions candidates comme constituant une entité nommée. Cette approche permet de lever la plupart des ambiguités et de reconnaître que "la ville lumière" fait allusion à Paris.

Les principaux "chantiers" de la reconnaissance des entités nommées

La désambiguation des noms de personne

Jacques Martin est à la fois le dessinateur d'Alix, et le nom d'un célèbre comédien et animateur de télévision. C'est aussi un entraineur de Hockey célébrissime au Canada. Il est donc indispensable de repérer dans un document de quel Jacques Martin on parle, ce qui permet soit de ne renvoyer que les pages qui parlent du dessinateur, soit d'afficher clairement dans la page de résultat de quel Jacques Martin parle tel ou tel document (solution choisie par le moteur de recherche Cuil par exemple). L'étude du contexte permet en général de résoudre ces problèmes d'homonymie. On trouve ces techniques également sur les moteurs Spock.com et Zoominfo.com



La page de résultats de Cuil sur Jacques Martin

L'identification des descriptions d'entités nommées

Il est important de repérer tous les cas d'expressions ne contenant pas de noms propres, mais utilisés à titre de périphrases pour décrire une entité nommée. Comme par exemple pour Paris : *la capitale, la ville lumière, la capitale de la mode*, ou plus compliqué pour Prince : l*ove symbol* ou *the Artist Formerly Known As Prince* !

La traduction des entités nommées

Contrairement à ce que l'on pourrait penser, de nombreux noms propres ne sont pas du tout invariants en passant d'une langue à l'autre. Pékin / Beijing, Aix La Chapelle / Aachen, Guillaume le Conquérant / William the Conqueror. Si l'on a affaire au cas de Prince, "l'artiste connu autrefois sous le nom de Prince" à la place de *Artist Formerly Known As Prince* !

C'est l'un des problèmes les plus délicats, car de mauvaises traductions produisent des résultats souvent ridicules ou absurdes, et dans un moteur de recherche renvoient franchement n'importe quoi ou rien du tout. Une part importante des erreurs commises par les outils de traduction automatique les plus performants viennent aujourd'hui d'erreurs de traduction sur les entités nommées.

L'identification des composants de noms de personnes

George W. Bush, *Pdt George Bush*, et *Président Bush* désignent la même personne, alors que *Laura Bush* est une autre personne. Pour parvenir à identifier ces cas, il est important de repérer au sein des groupes lexicaux représentant des entités nommées, les "composants" pouvant caractériser ces noms, comme des titres (*Dr, Pr, Jr*) ou des prénoms. Ces outils sont souvent intégrés en tant que "préprocesseurs" dans les systèmes de reconnaissance d'entités nommées. Dans le même style, on peut également citer la gestion des sigles et des acronymes.

La détection et la catégorisation de nouvelles entités nommées

Qui connaissait Cindy Sander avant qu'elle ne défraie la chronique en étant évincée d'une célèbre émission de télévision ?

Le web génère régulièrement des cas de noms de lieux ou de personnes, jamais cités dans aucune page auparavant, et qui, en raison d'une actualité, apparaissent en quelques heures sur des centaines, des milliers, des centaines de milliers de nouvelles pages web !

Il peut donc être intéressant de savoir détecter ces nouvelles entités nommées, surtout si elles sont génératrices de recherches sur les moteurs, et de savoir en plus les catégoriser. Ainsi, il sera intéressant de classer Cindy Sander dans les "célébrités", dans les "chanteuses", voire dans les participants dans les émissions de "télé-réalité". Jean Véronis a ainsi révélé sur son blog (http://aixtal.blogspot.com/2008/02/wikio-portail-dactualits-intelligent.html) que le site Wikio utilisait ce type d'approche.

Les avancées de Google, Yahoo et MSN sur la reconnaissance des entités nommées

Pour un moteur de recherche, savoir utiliser la reconnaissance des entités nommées permet de faire de grands progrès en matière de pertinence... L'une des premières utilisations est de pouvoir annoter les documents contenus dans l'index du moteur en procédant à de l'extraction d'informations au coeur du texte brut de ces pages. L'index ainsi constitué permet ainsi de sortir plus facilement, sur la requête *Tony Parker basketteur*, toutes les pages qui parlent de ce Tony Parker-là, même si le terme *basketteur* ne figure pas dans la page, même si la page contient son surnom (*TeePee*) au lieu de son nom, et en évitant de sortir en premier les pages sur *Tony Parker le plombier du Bronx*.

Les processus d'indexation modernes dans les moteurs utilisent de plus en plus ces systèmes de "taggage" qui ajoutent de l'information structurée dans les documents, selon un procédé assez proche de ce qui est prévu dans le web sémantique.

Ensuite ces outils sont très utilisés dans les systèmes de recherche universelle que l'on retrouve dans la plupart des grands moteurs. Par exemple, la recherche "locale" est un concept qui ne peut pas fonctionner sans exploiter les techniques de NER. Pour reconnaître que Vergèze est une ville et non une personne, il faut déjà avoir constitué des bases de villes à partir des pages du web (ou de bases de noms de villes existantes), et savoir reconnaître l'entité nommée *Vergèze* aussi bien dans la requête que dans les documents indexés.

Patrick Pantel de Yahoo a décrit dans une conférence donnée le 15 novembre 2008 (*NSF Symposium on Semantic Knowledge: Discovery, Organization and Use*, november 15, 2008) comment Yahoo exploite de manière combinée la reconnaissance des entités nommées, et la reconnaissances des intentions des utilisateurs, pour produire des snippets intelligents et d'autres enrichissements de l'interface utilisateur. Les trois approches combinées sont les suivantes :
- *Information Extraction: Entity detection and salience; attribute detection*
- *Content Analysis: Text classification, aboutness, information fusion*
- *Query Intent Modeling : Entity detection, intent/task understanding*

Google utilise également ces approches pour parvenir à adapter la présentation de ses résultats de recherche universelle (Books, Video, Images, News, Maps, Images) en fonction de la nature des entités nommées présentes dans les requêtes.

Les approches innovantes inventées par les moteurs de recherche

Les moteurs de recherches généralistes (Google, Yahoo ou Live...) doivent utiliser impérativement des approches universelles (qui fonctionnent dans toutes les langues et dans tous les contextes) et les plus automatiques possibles.

Il est apparu assez vite que pour faire un système de reconnaissance des entités nommées efficace, l'intervention à différents stades d'évaluateurs humains, ou de personnes chargées d'alimenter le système en informations était indispensable. Pour des outils très spécialisés, ce n'est pas un problème majeur, mais pour un moteur de recherche comme Google, cela crée

des coûts insupportables.

Les algorithmes d'apprentissage automatique au service de la reconnaissance des entités nommées

De nombreux travaux ont été consacrés à partir de 2004/2005 à la création d'algorithmes capables, à l'aide de jeux de tests et de quelques corrections "manuelles", d'apprendre automatiquement à reconnaître des entités nommées et à les catégoriser. Compte tenu de l'importance de ces approches pour ces moteurs, on peut noter que de très nombreuses publications sur ce sujet émanent des laboratoires de Yahoo, Live/MSN et de Google...

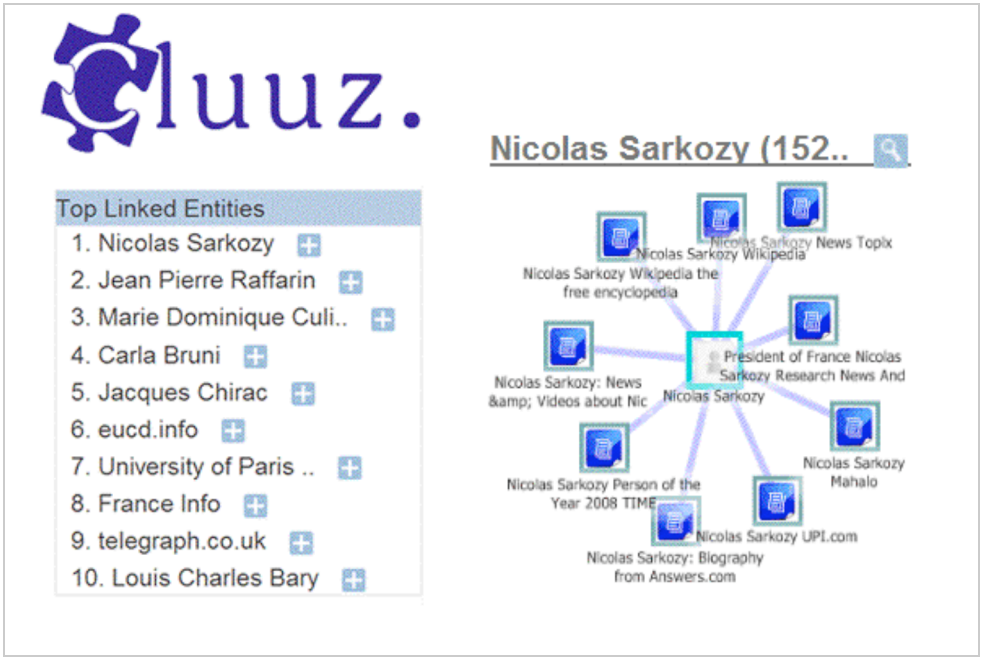
Ces travaux portent sur des systèmes d'apprentissage non supervisés ou légèrement supervisés (*supervised learning, semi-supervised leraning, lightly supervised learning*) .

L'identification automatique des concepts regroupant des entités nommées

C'est l'un des défis les plus importants pour les chercheurs qui travaillent sur la reconnaissance des entités nommées. C'est aussi le champ de recherche qui intéresse le plus des moteurs comme Google ou Yahoo...

Nous l'avons déjà signalé au début de cet article : les catégorisations "simplistes" ("villes", "personnes", "organisations") ne permettent pas de classer toutes les entités nommées. Qui plus est un moteur de recherche a besoin dans la pratique de savoir catégoriser les entités nommées de manière beaucoup plus fine, et même de les étiqueter avec des attributs supplémentaires permettant de les qualifier (*Tony Parker:personne,basketteur, Thierry Henry:personne,footballeur*)

L'approche classique est de faire travailler des documentalistes pour réussir à créer des catégorisations utilisables pour les besoins du moteur. Mais là encore, l'exercice peut se révéler trop coûteux, difficile à reproduire pour toutes les langues de la terre pour un moteur ayant une vocation universelle. Les chercheurs de Google en particulier ont développé des méthodes automatiques pour découvrir les "concepts" qui se trouvent derrière des groupes d'entités nommées, soit en s'aidant de corpus structurés (l'ontologie Wordnet ou l'encyclopédie en ligne Wikipedia), soit par des méthodes de pure linguistique statistique.



Cluuz, l'un des très nombreux moteurs "alternatifs" sortis depuis deux ans et utilisant intensivement les possibilités offertes par la reconnaissance des entités nommées.

Google et la reconnaissance des entités nommées par abstraction

Peter Norvig de Google avait révélé dans une conférence tenue en 2004 (*Google's Web 2 Demo and the UI Plunge, October 12, 2004, by John Battelle :* http://battellemedia.com/archives/000960.php) que Google préparait des innovations dans le domaine de la sémantique, et qu'ils envisageaient d'introduire une technologie qu'il appelait "la reconnaissance des entités par abstraction". Ces propos faisaient sans doute allusion aux travaux de Marius Pasca, un éminent chercheur récemment arrivé chez Google, et ceux de Thorsten Brants.

Marius Pasca, le spécialiste de la sémantique et des entités nommées chez Google...

Marius Pasca est un chercheur émérite de l'équipe de recherche de Google. Il a obtenu un Ph.D en Informatique de l'université méthodiste du Sud de Dalllas en décembre 2001, après un Doctorat es Sciences délivré par l'université Joseph Fourier de Grenoble en juin 1998. Il est l'auteur du livre "*Open-domain question answering from large text collections*", publié en avril 2003. Ses recherches actuelles portent sur l'extraction d'informations factuelles à partir de textes non structurés, et les fonctions avancées de correspondance pour la recherche d'information.

L'avenir de ces techniques dans les moteurs de recherche

Petit à petit, de manière assez silencieuse et assez peu remarquée, sauf par les spécialistes des TAL, les techniques de reconnaissance des entités nommées ont pris de l'importance dans le fonctionnement des moteurs leaders. Les avancées théoriques et les innovations pratiques des années 2003 à 2004 ont mis un peu de temps pour être exploitée dans les moteurs, car elles ont réclamé des changements parfois complexes de l'architecture des index des moteurs pour pouvoir être utilisés.

Ce délai de mise en place a été mis à profit par certains moteurs verticaux "alternatifs" qui ont abondamment exploité ces techniques pour montrer la supériorité de leurs algos sur celui de Google.

Mais ces derniers mois, une salve d'innovations exploitant largement les entités nommées, l'extraction d'information des documents et la détection de l'intention derrière la requête sont apparues sur Google, Live (MSN) et Yahoo. Il semble qu'une nouvelle architecture du moteur Google permette à ce dernier de multiplier ces derniers temps l'intégration d'applications de la NER dans son fonctionnement. Différents progrès de ces techniques laissent présager la multiplication d'applications nouvelles dans les années et les mois à venir. Certaines apparaissent déjà dans Google Labs, d'autres sont déjà de manière embryonnaires dans les pages de résultats, d'autres encore ne sont pas implémentées pour l'instant.

Ces innovations concernent six domaines principaux :

- **la détection d'évènements datés** (conjonction d'entités nommées de type date heure en conjonction avec d'autres entités).

- **l'analyse de la variation des entités au fil du temps** (permettant d'organiser une "ligne de temps" autour d'une personne, d'un évènement).

- **les systèmes de questions réponses** : en extrayant des faits de textes non structurés, on peut imaginer donner directement la réponse à une question posée en langage naturel.

- **l'extraction d'information sémantique** : permet de renvoyer une liste d'entités nommées appartenant à une catégorie quand cette catégorie est demandée dans une requête. Ou de renvoyer des entités nommées cousines quand la requête contient une entité nommée. Cette approche améliore la pertinence des résultats (les innovations dans ce domaine sont issues des travaux de Marius Pasca).

- **la recherche locale** : savoir reconnaître que l'intention de l'utilisateur est une recherche dans un contexte géographique précis permet de fournir des résultats géolocalisés et d'améliorer grandement la pertinence des résultats. Cette approche utilise la NER, et vient de faire une apparition remarquable dans les pages de résultats de Google en avril 2008.

- **l'extraction de connaissances** : l'analyse des pages internet en grand nombre permet de découvrir des informations reliant certains éléments qui ne sont pas disponibles dans les documents isolés. Ces informations peuvent ensuite servir à enrichir d'attributs et de relations les entités reconnues dans ces pages isolées.

On voit que le domaine de la reconnaissance des entités nommées est en pleine effervescence. Gageons que d'autres applications impressionnantes peuvent faire leur apparition dans les mois ou les années qui viennent. Et j'espère que cet article vous permettra de reconnaître ... la reconnaissance des entités nommées à l'oeuvre derrière ces applications.

Bibliographie

M. Pasca, "*Acquisition of Categorized Named Entities for Web Search*", inProceedings of the 13th ACM International Conference on Information and Knowledge Management(CIKM 2004), ACM Press, Washington, D.C., USA, 8 -13 November 2004, pp 137 –145.

Finding Cars, Goddesses and Enzymes: Parametrizable Acquisition of Labeled Instances for Open-Domain Information Extraction
Benjamin Van Durme Marius Pasca

The Role of Documents vs. Queries in Extracting Class Attributes from Text
Marius Pasca Google Inc. / Benjamin Van Durme University of Rochester / Nikesh Garera Johns Hopkins University

What You Seek isWhat You Get: Extraction of Class Attributes from Query Logs
Marius Pasca Google Inc. / Benjamin Van Durme University of Rochester

Outclassing Wikipedia in Open-Domain Information Extraction:
Weakly-Supervised Acquisition of Attributes over Conceptual Hierarchies, Marius Pasca

A Context Pattern Induction Method for Named Entity Extraction
Partha Pratim Talukdar CIS Department / Thorsten Brants Google, Inc. / Mark Liberman Fernando Pereira CIS Department

Vers une double annotation des Entités Nommées : Maud Ehrmann— Guillaume Jacquet Centre de Recherche Xerox de Grenoble

Maud Ehrmann - Présentation / Conférence IDL 17 Oct 2008 - Université Stendhal

Poibeau T., "Deconstructing Harry , une évaluation des systèmes de repérage des entités nommées", Revue de l'électricité et de l'électronique, EDP Sciences, 2001.

Poibeau T., "*Sur le statut référentiel des Entités Nommées*", Actes de la conférence Traitement Automatique des Langues Naturelles, Dourdan, France, Atala, 2005.

Of Search and Semantics, Patrick Pantel NSF Symposium on Semantic Knowledge Discovery, Organization and Use November 15, 2008

Nadeau, David (2007) : *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. [Thesis]

Wang, Lee, Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.-Y. and Li, Y. (2005) *Detecting Dominant Locations from Search Queries*. Proc. International ACM SIGIR Conference.

Brevets intéressants

Implicit name searching
Invented by Rajat Mukherjee, Irfan Presswala, and Kalpana Ravinarayanan
Assigned to Yahoo
US Patent Application 20080228720
Published September 18, 2008
Filed March 14, 2007
USPTO

Disambiguation of Named Entities
Inventors: Razvan Constantin Bunescu and Alexandru Marius Pasca
US Patent Application 20070233656
Published October 4, 2007
Filed: June 29, 2006
USPTO

Philippe Yonnet, Directeur Technique @Position (http://www.aposition.com) et président de l'association SEO Camp (http://www.seo-camp.org/)

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2009/05/la-reconnaissance-des-entites-nommees.html