Les entités nommées en TAL

Anthony BETREMIEUX
Institut National des Langues et Civilisations Orientales
ermaion@gmail.com

Contenu

Présentation de ce qu'on appelle les entités nommées par une brêve définition. On abordera les applications faites en TAL de ces entités nommées (principalement l'extraction d'informations) et le rôle grandissant de ces entités dans le travail des moteurs de recherche.

1 Définition

Une entité nommée est une appellation générique utilisée pour la catégorisation d'un certain nombres d'objets textuels rencontrés dans un document.
Cette appellation recouvre par exemple les noms de personnes, des lieux ou des organismes (raisons sociales). Par extension, on y rajoute les adresses mails, les numéros de téléphone, les codes postaux, et toute information de ce genre que l'on peut y relier.

Certaines bibliothèques reconnaissent jusqu'à près d'une trentaine d'entités nommées:

* Personne
* Fonction
    - politique
    - militaire
    - administrative
    - religieuse
    - aristocratique
* Date et heure
    - date absolue
    - date relative
    - heure
* Lieu
    - adresse
    - téléphone
    - adresse électronique
* Organisation
* Production humaine
    - moyen de transport
    - récompense
    - œuvre artistique
    - documentaire
* Montant
    - âge
    - durée
    - température
    - longueur
    - surface
    - volume
    - poids
    - vitesse
    - prix
    - autre montant

A titre d'exemple, on pourrait donner le texte qui suit, étiqueté par un système de reconnaissance d'entités nommées utilisé lors de la campagne d'évaluation MUC:

Henri a acheté 300 actions de la société AMD en 2006

<ENAMEX TYPE="PERSON">*Henri****</ENAMEX>*** *a acheté* ***<NUMEX TYPE="QUANTITY">****300****</NUMEX>*** *actions de la société* ***<ENAMEX TYPE="ORGANIZATION">****AMD****</ENAMEX>*** *en* ***<TIMEX TYPE="DATE">****2006****</TIMEX>****.*

Une rapide analyse de l'étiquetage des entités nommées du texte ci-dessus révèle une des manières de réaliser cet étiquetage, et l'on voit dans cet exemple que les entités nommées sont implémentées en XML.

2 Implémentation XML des entités nommées

L'implémentation se fait généralement en XML.
XML (eXtensible Markup Language) est un développement de la norme internationale SGML avec les apports du langage hypertexte de HTML.
Mais, à la différence de HTML, XML ne mélange jamais la mise en forme de la structure d'un document. A ce titre, XML est très souvent utilisé pour l'encodage de ressources, comme les méta-données ou donc les entités nommées.

XML présente plusieurs avantages:

● il est indépendant des logiciels et utilise
UNICODE, permettant une grande interopérabilité et une utilisation dans de multiples environnements.

● Il permet une structuration logique du contenu du document.

● Existence de modèles standards et partageables de documents, DTD et, dernièrement, le schéma XML.

● Le processus de transformation XSLT permet d'associer une feuille de style XSL, permettant une utilisation multiple des ressources XML.

● XML apparaît comme une simplification de l'implémentation sous SGML, dont il reprend seulement 10% des éléments. La structuration XML des métadonnées permet donc une plus grande pérennité du contenu, une meilleure accessibilité, une interopérabilté facilitant une conversion pour d'autres utilisations comme la partage de données (alimentation de portails et catalogues, d'archives, analyses statistiques).

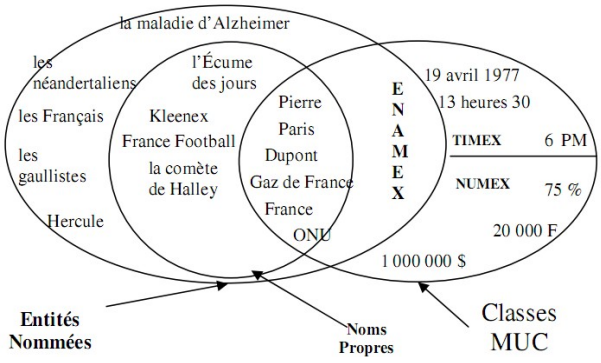
3 Les différents types d'entités nommées

Une étude du texte donné en exemple pour l'étiquetage d'entités nommées nous révèle que l'étiquetage procède selon les normes XML, c'est-à-dire avec des éléments, des attributs et des valeurs. Les éléments sont au nombre de trois (trois types définis par l'UMC (Message Understanding Conference):

- NUMEX (noms propres)
- TIMEX (expressions temporelles)
- NUMEX (expressions numériques)

+ ENAMEX: noms de personnes, noms de villes
+ TIMEX : date, heure
+ NUMEX : montants financiers, pourcentages

La notion d'entité nommée inclut donc les noms propres, mais aussi les gentilés, les personnages de légendes, les maladies ou les drogues qui ne sont pas toujours considérés comme des noms propres. La figure ci-dessous illustre la notion d'entité nommée.



4 les campagnes d'évaluation (un exemple: le MUC)

Le principe de ces campagnes est de fournir un corpus d'entraînement pour adapter le système

à la tâche d'étiquetage, et un corpus de test pour mesurer ses performances. Dans ces compagnes, les systèmes obtiennent régulièrement des scores supérieurs à 90% (de l'ordre de 95% lors des campagnes récentes), alors que les annotateurs humains obtiennent des scores supérieurs ou proches de 97%.
Ces scores sont certes à relativiser, puisque dans des conditions ouvertes (à savoir n'importe quel document fourni à un étiqueteur sans apprentissage), les meilleurs systèmes sont rarement au-dessus de 50% de performances.

Le Message Understanding Conferences (MUC) a été initié et financé par le DARPA (Agence de Projets de Recherche Avancée de la Défense) pour encourager le développement de nouvelles et de meilleures méthodes d'extraction d'informations.
Ce n'est que pour la première conférence (MUC-1) que l'on disposait d'un libre choix du format de sortie pour les informations extraites. A partir de la deuxième conférence, les formats de sortie par lequel les systèmes étaient evalués étaient prédéfinis à l'avance.
A la sixième conférence (MUC-6) tla question de la reconnaissance d'entités nommées et de coréférences a été ajoutée. Pour les entités nommées toutes les phrases étaient supposées être marquées en fonction des personnes, des lieux, du temps ou d'une quantité.

Les sujets et les textes sources qui faisaient l'objet d'un traitement ont révélé une lente évolution depuis des thèmes militairs vers des thèmes civils, reflétant ainsi les changements intervenus dans les intérêts commerciaux du marché à propos de l'extraction d'information. Ce qui nous donne l'occasion de porter notre attention sur l'une des applications de la reconnaissance des entités nommées: l'extraction d'informations.

5 Extraction d'information et Entités nommées

L'extraction des entités nommées est une tâche très classique en KM (Knowledge Management). Cette extraction consiste à "baliser" des documents textes pour l'identification et le catégorisation des mots dans des entités nommées correspondantes.

La reconnaissance, à partir de textes écrits, des entités nommées est actuellement la tâche d'extraction d'information obtenant les meilleures performances avec des taux combinés de précision et de rappel comparables à ceux des humains (de l'ordre de 0,90, sur des dépêches journalistiques par exemple).
Généralement, deux grandes approches prévalent pour leur identification : une approche linguistique (de surface), et une approche probabiliste.

5.1. une approche linguistique

Les approches linguistiques sont basées sur des règles génériques, écrites à la main, se fondant sur des informations lexico-syntaxiques du texte. Des règles de grammaire utilisent des marqueurs lexicaux (ex. Mr pour Mister ou Inc. pour Incorporated), des dictionnaires de noms propres et des dictionnaires de la langue générale (essentiellement pour repérer les mots inconnus) sont utilisés pour repérer et typer les syntagmes intéressants
Ces systèmes sont particulièrement performants sur des textes bien écrits (notamment des textes journalistiques) mais leur performance diminue considérablement sur des textes bruités (par exemple ceux issus d'une-transcription orale).

5.2. une approche probabiliste

Les approches statistiques sont fondées sur un mécanisme d'apprentissage à partir de textes étiquetés manuellement. Le résultat de l'apprentissage est sauvegardé pour être utilisé

dans le module de reconnaissance. Ces systèmes, robustes pour des textes bruités (comme par exemple la grande majorité des systèmes dédiés à l'oral, qui adoptent une telle approche), nécessitent en contre-partie un important volume de textes étiquetés dans la phase d'apprentissage.

5.3. une combinaison des deux approches

Récemment des approches hybrides ont vu le jour, tirant parti des avantages respectifs des méthodes statistique et linguistique. Avec de tels systèmes, un ensemble de règles est généralement appris automatiquement puis révisé par un expert. L'approche inverse a aussi été testée : élaboration par un linguiste d'un ensemble de règles de base puis étendu (semi-) automatiquement par un moteur d'inférence permettant d'obtenir progressivement une couverture optimale du corpus

6 Moteurs de recherche et Entités Nommées: vers une notion de Web sémantique

On assiste à une profusion des moteurs de recherche. Si certains d'entre eux ne présentent qu'un intérêt limité (notamment ceux qui reposent sur la linguistique statistique sans aucune touche de sémantique), d'autres (comme par exemple iSeek ou Evri), font partie de ces nouveaux moteurx de recherche qui basent leur originalité sur l'exploration des entités nommées et l'analyse des relations entre elles sur fond d'AI (Intelligence Artificielle).

Tous ces moteurs de recherche privilégient une approche axée sur la sémantique appliquée, c'est-à-dire qu'ils utilisent des outils sémantiques, afin de mieux se différencier des leaders du marché (par exemple Yahoo, Google, etc.) qui ne privilégient pas une telle approche (ou alors de manière voilée, par exemple les publicités de Google Adsense...).

Parmi ces moteurs de recherche orientés sémantique, citons Webfountain, une architecture proposée par IBM permettant la réalisation de puissants moteur de recherche (avec les machines, les unités de stockage, et les logiciels de crawl, de data-mining ainsi que le logiciel de l'outil de recherche).

Sur l'aspect sémantique, l'un des apports majeurs d'IBM est la mise en oeuvre d'un outil très souple de balisage sémantique, appelé SEMTAG, qui est un fantastique outils en data mining (extraction d'informations) par l'identification des entités nommées qu'il effectue.

7 Les difficultés d'une approche sémantique: la polysémie de certaines entités nommées

La polysémie de certaines entités nommées vient compliquer passablement l'analyse automatique. L'opération qui consiste à catégoriser une entitée repose sur le postulat de leur référentialité. Mais une entité nommée peut référer à plusieurs classes.
Les dates se confondent ainsi fréquemment
avec des événements :

Le 11 septembre 2001 a représenté un tournant dans l'histoire américaine. (Elie Wiesel,
site www.france-amerique.com)

Il est parfois difficile de classer les noms d'organisations, qui peuvent être catégorisés comme institution, communauté d'individus ou encore bâtiment.

Le journal télévisé a eu lieu hier en direct de l'ONU.

L'ONU était en grève hier.

L'ONU a fêté ses 50 ans.

L'ONU n'acceptera pas une attaque frontale de l'Irak (forum du Monde)

Un nom de personne peut référer à une œuvre, à un objet ou à tout autre élément ayant un lien direct ou non avec la personne.

J'ai tout Chirac sur l'étagère

A travers ces exemples rapides on se rend compte que les entités nommées ne diffèrent pas fondamentalement des autres unités linguistiques. Un des exemples parmi les plus connus est celui du du Prix Goncourt.
Le Prix Goncourt renvoie à plusieurs sens différents, suivant qu'il s'agisse du prix, de sa valeur monétaire, du livre qui a obtenu le prix, de l'institution ou encore de l'auteur.

Il résulte de ce qui précède qu'il peut être nécessaire d'avoir une modélisation à base de structures de traits permettant de rendre explicite les différentes désignations en discours, en unifiant le mode de représentation des connaissances linguistiques et des connaissances sur le monde.

Indications bibliographiques

http://fr.wikipedia.org/wiki/Entit%C3%A9s_nomm%C3%A9es

http://www.synapse-fr.com/API/Extraction_Entites_Nommees.htm

http://www.webmaster-hub.com/publication/imprimer115.html

http://www.technolangue.net/imprimer.php3?id_article=295

http://www.hal.archives-ouvertes.fr/docs/00/03/74/98/PDF/taln05-poibeau.pdf

http://www.univ-tlse2.fr/erss/textes/publications/CDG/25/CG25-7-Daille.pdf