ANNOTATION SÉMANTIQUE

Entités nommées

EXTRACTION D'INFORMATIONS

- Enjeux actuels : Arriver à un traitement automatique de grandes masses de données (Big Data)
- se fixe comme objectif d'*extraire* de textes des informations factuelles précises et/ou d'identifier les actants de certaines situations (des attentats aux interactions géniques), cette situation étant décrite par un formulaire à instancier avec les informations extraites du texte.
- exemple
  - les textes de petites annonces de vente de voitures, rédigées librement.
  - qui vend quelle type de voiture, de quel kilométrage, à quel prix, etc.
- *wrapper* (terme anglais qui signifie "envelopper") :
  - un programme capable de remplir automatiquement les valeurs de ces champs à partir du texte initial de la petite annonce.
  - est nécessairement spécialisé dans le traitement d'un certain type de textes

WRAPPER

- les informations disponibles au wrapper
  - la liste des champs à extraire
  - Les ressources linguistiques utiles
- Techniques pour définir un « wrapper »
  - des automates ou des expressions régulières qui repèrent les environnements possibles où peut apparaître l'information visée
  - l'apprentissage automatique de wrappers à partir d'exemples de textes d'où ont été extraits des données factuelles

EXTRACTION D'INFORMATIONS

- Applications du TAL en relation avec l'extraction d'informations :

  - La recherche d'informations :
    - Manière de retrouver des informations dans un corpus
  - Indexation automatique
  - La Fouille de texte ou Text Minning :
    - ensemble des traitements informatiques destinés à extraire des connaissances selon des critères définis dans des textes produits par des humains.
  - Le résumé automatique

EXTRACTION : PRINCIPES

- **l'extraction "libre" :**

l'information à extraire n'est pas définie auparavant. L'objectif est de décrire et représenter le contenu global des documents (l'extraction de mots ou de phrases "représentatifs" du texte). L'étape de l'extraction doit être suivie par celle de filtrage d'information qui permet de nettoyer les résultats des mots non représentatifs ;

- **l'extraction "normalisée" :**

l'objectif est de repérer les mots et de les extraire selon des catégories prédéfinies (personne, organisation, lieu, produit, etc.).

LES ENTITÉS NOMMÉES

LES ENTITÉS NOMMÉES

… noms propres au sens classique, noms propres dans un sens élargi mais aussi expressions de temps et de quantité

(MUC-A, Chinchor 1998, Ehrmann, 2008)

… des éléments informationnels pertinents dont on parle et qui jouent un rôle dans la description d'un événement, d'un fait

(Nouvel *et al.*, 2015, 2013)

…toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus

… tous les éléments du langage qui font référence à une entité unique et concrète, appartenant à un domaine spécifique (ie. Humain, économique, géographique, etc.)

(Ehrmann 2008)

CAMPAGNES D'ÉVALUATIONS

CAMPAGNES D'EVALUATION

Pour comparer les résultats obtenus par différentes équipes de Tal,

- l'idée est de définir des annotations à insérer dans un corpus et d'évaluer les résultats de chaque équipe, à la fois en terme de précision (a-t-on annoté correctement une entité ?) et de rappel (en a-t-on oublié certaines ?).

Une campagne d'évaluation propose

- des annotations à insérer dans un corpus sur lequel tous les systèmes participants seront évalués en terme de rappel et précision (et/ou F-mesure).
- Un guide d'annotation précise le résultat attendu.
- Un corpus d'entraînement et de test, notamment pour les systèmes à base d'apprentissage.
- L'annotation manuelle est elle-même contrôlée par des mesures d'accord inter-annotateurs (Artstein et Poesio, 2008) (Fort et al., 2012).

MUC (MESSAGE UNDERSTANDING CONFERENCE)

- Le but de ces conférences
  - Evaluer les systèmes d'extraction d'informations.
- La définition est donnée par (Chinchor, 1997) pour la conférence Muc-7 :
  - On the level of entity extraction, Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts.

D'AUTRES CAMPAGNES

- Met (Merchant et al., 1996) pour le japonais, le chinois et l'espagnol,
- Irex (Sekine et Eriguchi, 2000) pour le japonais,
- Harem (Santos et al., 2006) pour le portugais
- Language-Independent Named Entity Recognition - des conférences CoNLL-2002 et CoNLL-2003, entièrement basées sur l'apprentissage.
- *Ester-2, 2008,* utilisait un guide d'annotation (Ester-2, 2007), testé lors de la campagne Ester-1, puis revu et disponible en ligne.
- *Etape, 2012,* s'est appuyée sur le guide d'annotation défini dans le cadre du projet Quaero (Rosset et al., 2011)

VISION GLOBALE

- Deux directions principales
  - des travaux dans le domaine “ général ”
    - la poursuite de la tâche définie par MUC pour d’autres langues que l’anglais,
    - un jeu de catégories plus ou moins revisité
    - annotation des entités dans des corpus de nature journalistique essentiellement (les campagnes d’évaluation MET, IREX, CoNNL, ACE, ESTER et HAREM)
  - des travaux dans des domaines dits “ de spécialité ”
    - la reconnaissance d’entités dans les domaines de la médecine, de la chimie ou de la microbiologie.
    - proposition de reconnaissance des noms de gènes, de protéines, etc. dans de la littérature spécialisée, lors des campagnes JNLPBA (Kim et al., 2004) et BioCreAtIvE (Hirschman et al., 2005).

CAMPAGNES EN FRANÇAIS

- Deux campagnes sur des corpus oraux transcrits, soit manuellement, soit automatiquement ont été organisées sous l'égide de l'association francophone de communication parlée (AFCP)

  - *Ester-2, 2008,* utilisait un guide d'annotation (Ester-2, 2007), testé lors de la campagne Ester-1, puis revu et disponible en ligne.

  - *Etape, 2012,* s'est appuyée sur le guide d'annotation défini dans le cadre du projet Quaero (Rosset et al., 2011)

Campagne	**Nombre d'étiquettes**	**Nombre de pages**
Muc-7	7	23
Ester-2	37	25
Etape	54	86

CAMPAGNE ESTER

- Menée par l'Association Francophone de la Communication Parlée (AFCP) et le Centre d'Expertise Parisien de la Délégation Générale de l'Armement (ELDA)
- Deux phases :
  - ESTER 1 : janvier 2005
  - ESTER 2 : janvier 2008 à avril 2009
- But :
  - Mesurer les performances des systèmes de transcriptions automatique d'émissions radiophoniques
  - Annoter manuellement en entités nommées (EN) le corpus en vue du développement de systèmes automatiques de détections des EN

CONVENTION D'ANNOTATION ESTER 2

- A chaque type d'entités nommées correspond une étiquette.

<pers.hum> L'Orléanaise Jeanne </pers.hum>

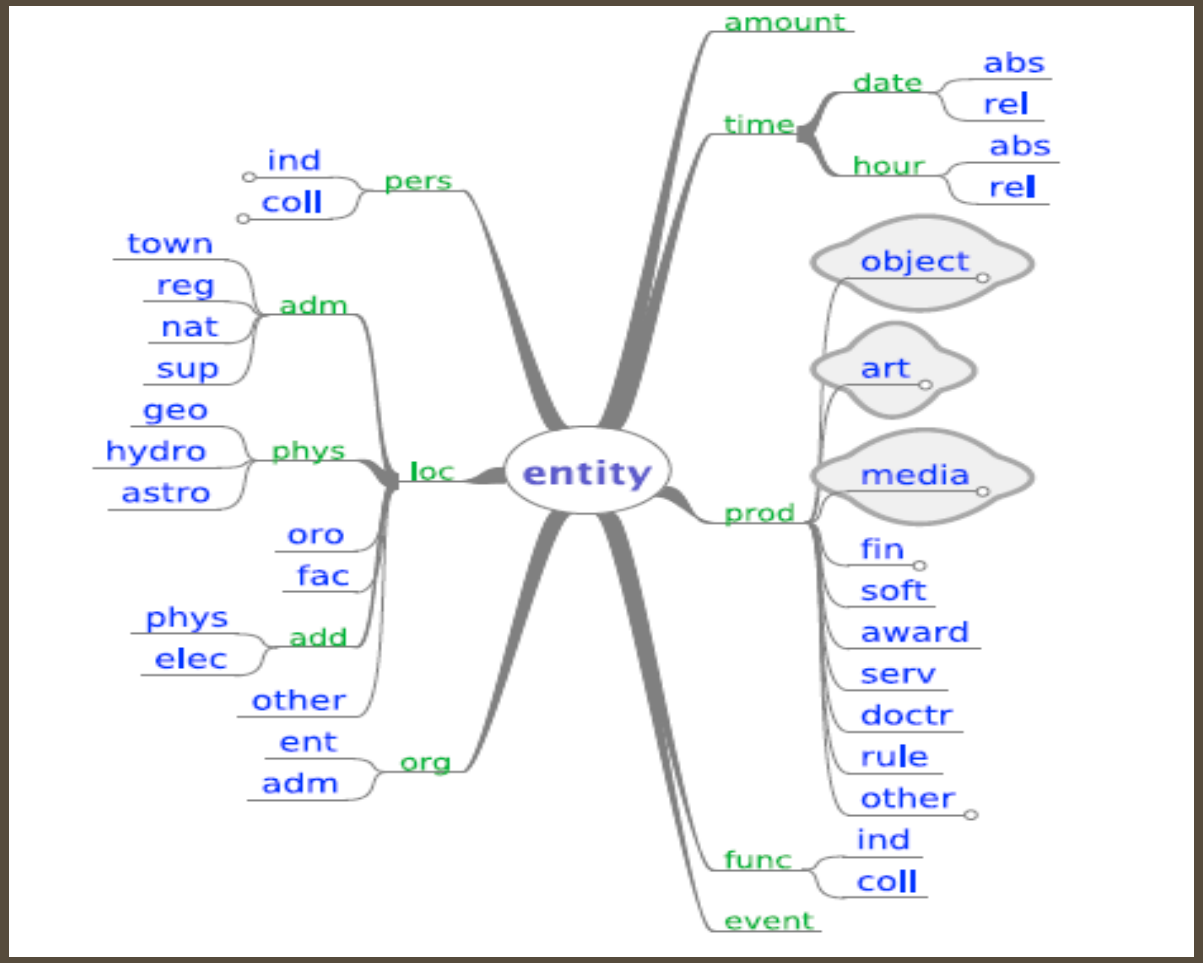
Convention 1.2	Pour les différentes catégories et sous-catégories identifiées, les valeurs xxx et yyy prennent les valeurs suivantes :
- personne ▪ humain réel ou fictif ▪ animal réel ou fictif	- pers ▪ pers.hum ▪ pers.anim
- fonction ▪ politique ▪ militaire ▪ administrative ▪ religieuse ▪ aristocratique	- fonc ▪ fonc.pol ▪ fonc.mil ▪ fonc.admi ▪ fonc.rel ▪ fonc.ari
- organisation ▪ politique ▪ éducative ▪ commerciale ▪ non commerciale ▪ média & divertissement ▪ géo-socio-administrative	- org ▪ org.pol ▪ org.edu ▪ org.com ▪ org.non-profit ▪ org.div ▪ org.gsp
- lieu ▪ géographique naturel ▪ région administrative ▪ axe de circulation ▪ adresse o adresse postale o téléphone et fax o adresse électronique ▪ construction humaine	- loc ▪ loc.geo ▪ loc.admi ▪ loc.line ▪ loc.addr o loc.addr.post o loc.addr.tel o loc.addr.elec ▪ loc.fac
- production humaine ▪ moyen de transport ▪ récompense ▪ œuvre artistique ▪ production documentaire	- prod ▪ prod.vehicule ▪ prod.award ▪ prod.art ▪ prod.doc
- date et heure ▪ date o date absolue o date relative ▪ heure	- time ▪ time.date o time.date.abs o time.date.rel ▪ time.hour
- montant ▪ âge ▪ durée ▪ température ▪ longueur ▪ surface et aire ▪ volume ▪ poids ▪ vitesse ▪ autre ▪ valeur monétaire	- amount ▪ amount.phy.age ▪ amount.phy.dur ▪ amount. phy.temp ▪ amount. phy.len ▪ amount.phy.area ▪ amount.phy.vol ▪ amount. phy.wei ▪ amount. phy.spd ▪ amount.phy.other ▪ amount.cur

CAMPAGNE ETAPE

- Menée AFCP (l'Association Francophone de la Communication Parlée) et l'ELDA (le Centre d'Expertise Parisien de la Délégation Générale de l'Armement (ELDA)
- De 2011 à 2012
- But :
  - Mesurer les performances des technologies vocales appliquées à l'analyse des flux télévisés en langue française.
  - Annotation en EN Quaero

ANNOTATION DES ENTITÉS NOMMÉES

Le guide Quaero s'affirme *en cohérence avec la définition des entités nommées* de la campagne Ester-2. Il adopte :
– la définition des entités nommées de la thèse de Maud Ehrmann (2008) ;
– la définition des entités temporelles du projet Time ML (Pustejovsky, 2003);
– la taxinomie de la thèse de Mickaël Tran (Tran, 2006 ; Tran et Maurel, 2006) ;
– la hiérarchie des entités nommées étendues proposée par Satoshi Sekine (Sekine,2004, Nadeau and Sekine, 2007).
–étend l'annotation à des expressions construites autour de noms communs



L'annotation n'est plus uniquement centrée sur les entités, mais aussi sur les composants de ces entités

Les composants principaux sont <*name*> pour le nom de l'entité et <*kind*> ou <*qualifier*> pour les noms communs et les qualifieurs constituant l'entité étendue.

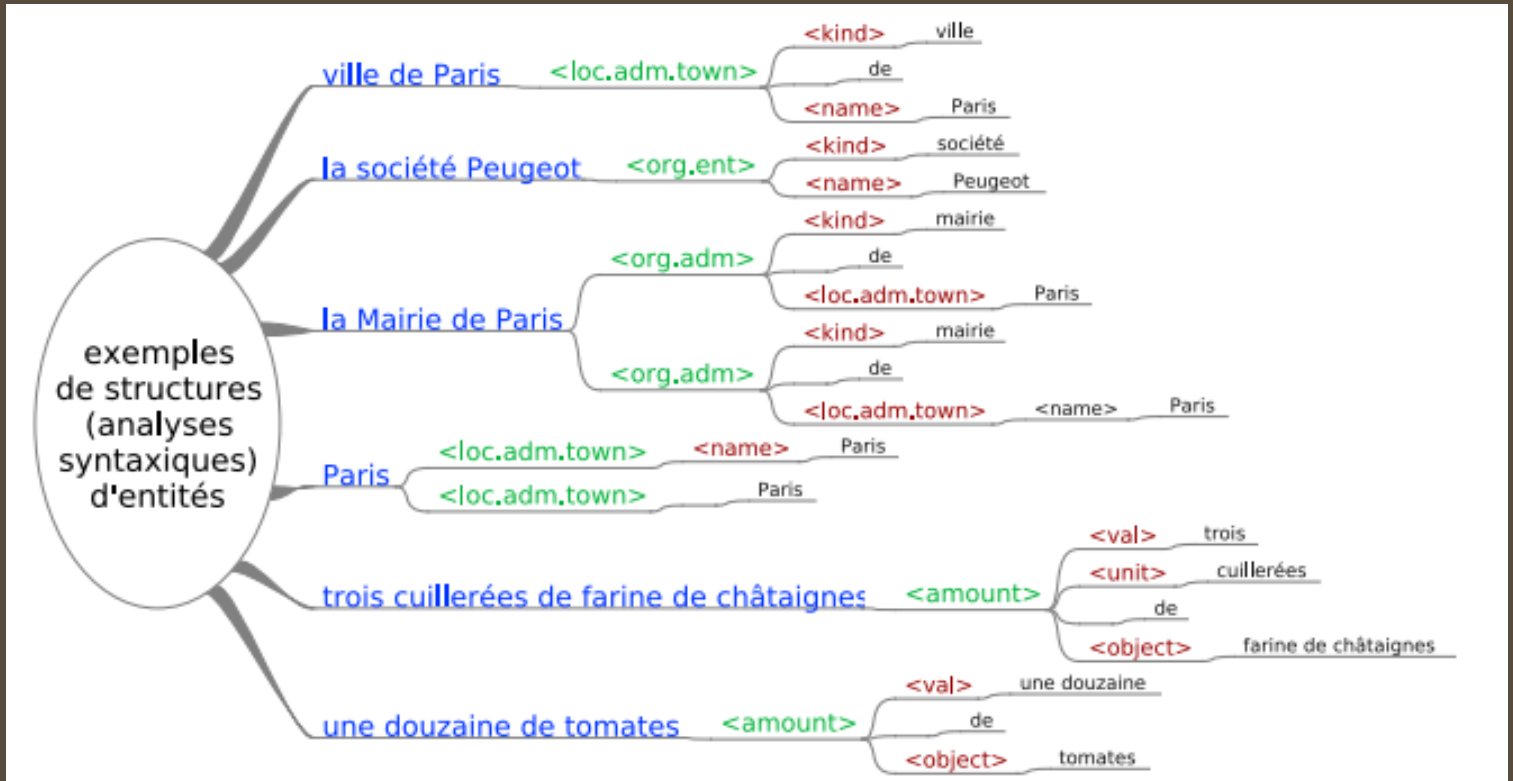


FIG. 1.3: Exemples d'entités.

CONVENTIONS D'ANNOTATIONS QUAERO

Utilisation de balises pour :

- Délimiter l'entité nommée
- Définir la nature de l'entité :
  - Personnes
  - Fonctions
  - Organisations
  - Lieux
  - Productions humaines
  - Informations de temps
  - Evénements

PERSONNES : PERS (<PERS.IND>, <PERS.COLL>)

Jeanne d'Arc

<pers.ind><name.first>Jeanne</name.first> <name.last>d'Arc</name.last></pers.ind>

les orléanais

les <pers.coll><demonym>orléanais</demonym></pers.coll>

Astérix et Obélix

<pers.ind>Astérix</pers.ind> et <pers.ind>Obélix</pers.ind>

Marie et Jean Dupont

<pers.coll><name.firstname>Marie</name.firstname> et <name.firstname>Jean</name.firstname>
<name.lastname>Dupont<name.lastname ></pers.coll>

FONCTION : FUNC (<FUNC.IND>, <FUNC.COLL>)

Le maire d'Orléans

Le <func.ind><kind>maire</kind> d'<loc.adm.town>Orléans</loc.adm.town></func.ind>

Le maire Olivier Carré

Le <func.ind><kind>maire</kind><func.ind>
<pers.ind><name>Olivier</name><name>Carré</name></pers.ind>

Les chercheurs du CNRS

Les <func.coll><kind>chercheur</kind> du <org.ent>CNRS</org.ent></func.coll>

ORGANISATIONS : ORG (<ORG.ENT>,<ORG.ADM>)

L'université d'Orléans

L'<org.ent><name>université</name> d'<loc.adm.town>Orléans</loc.adm.town></org.ent>

La mairie d'Orléans

La <org.adm><kind>mairie</kind> d'<loc.adm.town>Orléans</loc.adm.town></org.adm>

Le ministère de la culture

Le <org.adm><kind>ministère</kind> de la <name>culture</name></org.adm>

LIEUX : LOC

- Lieux administratifs : <loc.adm.town> – <loc.adm.reg> – <loc.adm.nat> – <loc.adm.sup>

  Orléans – Loiret/Centre – France – Europe
- Lieux physiques : <loc.phys.geo> – <loc.phys.hydro> – <loc.phys.astro>

  Mont Blanc – Loire – Lune
- Bâtiments : <loc.fac>

  Le Fnac – la cathédrale d'Orléans – l'Elysée
- Voies : <loc.oro>

  Autoroute A11 – RN20 – Place du Martroi – Place d'Arc ?
- Adresses : <loc.add.phys> et <loc.add.elec>
  - 10 rue de Tours

  <loc.add.phys><address-number>10</a-n> <kind>rue</rue> de <loc.adm.town>Tours</loc.adm.town><loc.add.phys>

  02 38 49 25 00 – 88.3 MHz – @Univ_Orleans

PRODUCTIONS HUMAINES : PROD

- Marque d'objet : <prod.objet>

  La Citroën C3 – Big Mac – Coca-Cola

- Œuvre artistique : <prod.art>

  Guernica – La Flûte Enchantée – Jurassic World

- Production médiatique : <prod.media>

  Le Monde – Secret Story – Le 19:45

- Produits financiers : <prod.fin>

  SMIC – le Dollar (a été réévalué) – PEL

- Logiciels : <prod.soft>

  Windows 10 – Facebook – système d'exploitation Unix

- Prix divers : <prod.award>

  Prix Nobel de la Paix – le César de la meilleur actrice

- Ligne de transport : <prod.serv>

  RER B – l'Intercité Paris-Orléans – Ligne A

- Doctrine : <prod.doctr>

  Le socialisme – le capitalisme – le christianisme

- Rule – Les lois : <prod.rule>

  Le Traité de Versailles

- Autres productions : prod.other Monopoly – Super Mario Kart

QUANTITÉ : AMOUNT

- Quantités : une valeur assortie d'une unité de mesure, suivie éventuellement d'un objet

22 bleus

<amount><val>22</val> <unit><pers.coll>bleus</pers></unit></val>

10 ans

J'ai <amount><val>10</val> <unit>ans</unit></val>

Quelques mètres

<amount><val>Quelques</val> <unit>mètres</unit></val>

- Durées :

une semaine entière

<amount><val>une</val> <unit>semaine</unit><qualifier>entière</qualifier></val>

pendant trois ans

<amount><qualifier>pendant</qualifier> <val>trois</val> <unit>ans</unit>

TEMPS : TIME

- Date : <time.date> – Doit être composé d'un jour de semaine, d'un jour (1 … 31), d'un mois, d'une saison, d'une année ou d'une référence temporelle (av/ap JC)

<time.date.abs> : jeudi 8 octobre 2018 – vers le 12 novembre

<time.date.rel> : jeudi prochain – la semaine dernière

- Heure : <time.hour>

<time.hour.abs> : quinze heure trente – 3 heures du matin

<time.hour.rel> : dans deux heures – entre 3 et 5 heures

EVÉNEMENTS : EVENT

- Aucune distinction entre les types d'événements pour plus de souplesse : les composants possibles d'un événement sont toutes les entités et tous les composants qui existent dans la convention. Tout doit être annoté.

Tour de France – Festival de Loire – Coupe du Monde de Rugby 2015

Fête Jeanne d'Arc

<event>Fête de <pers.ind><name.firstname>Jeanne</name.firstname><name.lastname>d'Arc</name.lastname></pers.ind></event>

LIENS UTILES

Conventions d'annotations

ESTER : http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf

QUAERO : http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

Outils d'annotation

CasEN : http://tln.li.univ-tours.fr/Tln_CasEN.html

Python : NLTK, SpaCy

Demo en ligne

http://apps.lattice.cnrs.fr/sem/

https://explosion.ai/demos/displacy-ent

https://cloud.gate.ac.uk/shopfront/displayItem/french-named-entity-recognizer

ANNOTATION DES ENTITÉS NOMMÉES : DIFFICULTÉS

Ambiguïté :

Orange, Christian Dior, Maison blanche

< ? ?>Yves Rocher</ ? ?> va s'installer à Vannes (pers/entreprise)

Métonymie (Figure de style par laquelle on exprime un concept au moyen d'un terme désignant un autre concept qui lui est uni par une relation nécessaire (cause et effet, inclusion, ressemblance, etc.) :

France en tant que lieu géographique, gouvernement ou équipe sportive

=> Quaero : lorsque le contexte ne permet pas de désambiguïser l'entité, un attribut : **class-set**

Paris

<entity class-set="loc.adm.town org.adm">

<name> Paris </name>

</entity>

ANNOTATION DES ENTITÉS NOMMÉES : DIFFICULTÉS

Nature du corpus : ex. Corpus du 19ème siècle

- Termes archaïques
- Civilités

  femme Lepin (corpus) / *madame Lepin* (aujourd'hui)

  fille Clavreuil (corpus) / *mademoiselle Clavreuil* (aujourd'hui)

- Abréviations : *vve* pour *veuve*
- Termes du domaine : *comre* pour *commissaire*

ANNOTATION DES ENTITÉS NOMMÉES : DIFFICULTÉS : DISFLUENCES

Hésitation : **j'ai une Peugeot euh 107**

j'ai une <prod.object> <org.ent> Peugeot </org.ent> euh <name> 107 </name> </prod.object>

Correction : **j'ai une Renault euh Peugeot 107**

j'ai une <prod.object> <org.ent> Renault </org.ent> </prod.object> euh
<prod.object> <org.ent> Peugeot </org.ent> <name> 107 </name> </prod.object>

ANNOTATION DES ENTITÉS NOMMÉES : DIFFICULTÉS : DISFLUENCES

Hésitation et correction : **tout au long euh euh de la semaine euh non de l'année**

<time.date.rel>
<time-modifier> tout au long </time-modifier>
euh euh de la
<kind> semaine </kind>
</time.date.rel>
euh non
<time.date.rel>
de l'
<kind> année </kind>
</time.date.rel>

Répétition : **Jean, Jean-Marie Le Pen**

<pers.ind>
<name_first> Jean </name_first>
</pers.ind>
,
<pers.ind>
<name_first> Jean-Marie </name_first>
<name_last> Le Pen </name_last>
</pers.ind>

ANNOTATION DES ENTITÉS NOMMÉES : DIFFICULTÉS

Coordinations : vallées de la Lorraine, de l’ Alsace, de la Bourgogne, de la Champagne

<loc.phys.geo>
<kind> vallées </kind>
de la
<name> Lorraine </name>
</loc.phys.geo>
,
<loc.phys.geo>de l’ <name>Alsace</name> </loc.phys.geo>
,
<loc.phys.geo>de la <name> Bourgogne</name> </loc.phys.geo>
,
<loc.phys.geo>de la <name> Champagne</name></loc.phys.geo>

ENTITÉS NOMMÉES : DÉFINITION : DIFFICULTÉS

- **EN sont des "mentions" qui renvoient à des "entités" du domaine, ces mentions pouvant relever de différentes catégories linguistiques : des noms propres ("Rabelais"), mais aussi les pronoms ("il"), et plus largement des descriptions définies ("le père de Gargantua").**

- Critères définitoires:
  - **Unicité référentielle** : un nom propre renvoie à une entité référentielle unique, même si cette unicité est contextelle.
  - **Autonomie référentielle** : permettent à elles seules l'identification du référent, tout au moins dans une situation de communication donnée
  - **Stabilité dénominative** : sont des dénominations stables, même s'il y a des variations (Angela Merkel/Mme Merkel/A. Merkel), elles sont plus régulières et moins nombreuses que pour les autres syntagmes nominaux
  - **Relativité référentielle** : l'interprétation se fait toujours relativement à un modèle du domaine, qui peut être implicite dans les cas simples (on suppose une connaissance partagée de ce qu'est une personne ou un pays) mais qui doit être explicité dès que la diversité des entités à prendre en compte s'accroît (il faut au moins une typologie pour les catégoriser).

(LDC, 2004)

NOM PROPRE VS NOM COMMUN

Critères de distinction :

- la majuscule,
- l'impossibilité de traduction,
- l'absence de l'article, l'incompatibilité avec des déterminants,
- la mono-référentialité,
- le manque de sens

NOM PROPRE VS NOM COMMUN

- des noms propres peuvent renvoyer vers plusieurs référents (*M. Dubois, Paul*)

- des noms propres peuvent renvoyer à la catégorie des référents

  "*Il y a souvent un Ernest Backes derrière les scoops. Un anonyme blessé, autodidacte, un temps favori des puissants, éjecté sans égards ensuite, qui règle ses comptes au nom d'un combat désintéressé pour la justice et la démocratie*"(p.23)

- des noms propres peuvent exister linguistiquement sans désigner des individus réels (*personnage mythologique*)

- des noms communs peuvent assurer une désignation unique

  - *lune,*
  - *le boucher vient tout à l'heure* (si l'on est dans un petit village qui n'a qu'un boucher)).

"Ce critère de l'unicité référentielle est donc lui aussi discutable, bien qu'en partie fondé. S'il correspond bien au fonctionnement du nom propre, il ne peut suffire à le définir ni à en délimiter la catégorie, ne serait-ce que parce que le nom commun y répond également dans certains cas, et que le nom propre y échappe dans d'autres cas" (Leroy 2004 : 24)

- les noms propres n'assurent pas seulement l'efficace désignation d'un référent du monde. En désignant quelqu'un par son prénom ou par son statut (*Madame*) on ajoute une information sur ses origines ou son statut civil.

EN ET APPLICATIONS

- REN
  - identifier les actants de certaines situations, cette situation étant décrite par un formulaire à instancier avec les informations extraites du texte.
- Systèmes de questions/réponses : quand est né Mozart ?
  - la question est tout d'abord envoyée à un système de *classification* pour être triée en fonction du type de réponse qu'elle requiert : ici, une réponse de type "date-de-naissance".
  - la question est par ailleurs analysée pour en extraire les mots clés significatifs : ici, "Mozart"
  - le ou les mots clés identifié(s) servent ensuite de requête à un système de *recherche d'information*, afin de concentrer la recherche sur des documents pertinents (ici : ceux qui parlent de Mozart).
  - les documents sélectionnés passent au crible d'une *extraction d'information* qui se focalise sur le type de données requis par le *type* de la question, tel qu'identifié lors de la première étape.
- Indexation automatique
  - Les EN sont utilisées comme "descripteurs"
- Aide à la lecture et à la navigation dans de gros volumes documentaires
  - On surligne les EN

EN ET APPLICATIONS

- Intégration de données : on s'appuie sur les EN référencées dans les bases de données pour établir des liens entre les différentes sources.
  - l'analyse des bases documentaires (suivi de thèmes, découverte de communautés de pratiques (Li & Liu, 2005)
  - l'articulation des documents avec d'autres sources de connaissances (bases de données, bases d'images, etc.) pour interroger les unes et les autres de manière homogène et naviguer facilement de l'une à l'autre (Dragos & Nazarenko, 2009).
- Anonymisation de documents
  - la ville qui a reçu la première bombe atomique, son père a fondé le plus grand cabinet ophtalmologique de la ville
  - Macron / Mbaye / Kanaan
- "Peupler" des ontologies
  - le modèle formel des ontologies impose de distinguer les entités du domaine, qui sont représentées comme des instances, des concepts ou classes auxquelles ces instances se rattachent.
  - La REN est alors utilisée pour enrichir la structure conceptuelle avec des instances de concepts ou de rôles (relations entre instances)

TECHNIQUES UTILISÉES

Méthodes symbolique

Fouille de texte : machine learning

Deep learning

https://www.youtube.com/watch?v=mcD2nEmKXUE (spacy)

https://www.youtube.com/watch?v=LFXsG7fueyk (NLTK, Python)

https://www.youtube.com/watch?v=MY9fs1Plh_o (évaluation Chritopher Manning)

https://www.youtube.com/watch?v=9qz1yEQlVhg (Spacy tutoriel)

https://www.youtube.com/watch?v=ili0JW4wylc (Deep learning : LSTM)

MÉTHODES SYMBOLIQUES :

Expressions régulières = combinaison de caractères littéraux et spéciaux permettant d'identifier une chaîne de caractères répondant à ce modèle.

MÉTHODES SYMBOLIQUES

Lexiques

- les listes “d’amorces” : des mots qui peuvent exprimer la catégorie recherchée :

société, compagnie, filiale, etc.

président, ministre, directeur, etc.

-Les dictionnaires

liste de prénoms, d’entreprises

- Les Patrons (spécifiques à la langue traitée): les règles de comportement syntaxique :

amorce + de + Groupe Nominal => société de service informatique, université de Paris Nanterre, etc.

amorce + prenom + NP => Président John Withmann

- Relations

Verbes => Spigadoro Inc. acquires largest European private label pasta company

-Marqueurs linguistique

ainsi, donc, en premier lieu => Phrases importantes

MÉTHODES SYMBOLIQUES :

limites :

- variation lexicale (synonymes, métaphore, etc.)
- l'emploi des termes anaphoriques
- on ne peut pas tout couvrir

=>

Approches statistiques : machine learning

EXPRESSIONS RÉGULIÈRES : MÉTACARACTÈRES

. remplace un caractère quelconque, désigne n'importe quel caractère

[ ] désigne des caractères compris dans un certain intervalle

196[89]=1968 ou 1969

12[.?]=12.ou12?

^ début de ligne

[^ opérateur d'exclusion : ne contenant pas les caractères suivants

$ fin de ligne

a$ = les lignes se terminant par le caractère a

- sépare deux caractères à l'intérieur des crochets, définit un intervalle de caractères

[0-9] couvre un chiffre quelconque

[a-z] couvre une minuscule non accentué quelconque

EXPRESSIONS RÉGULIÈRES

*	répétition de motif 0 ou n fois
+	1 ou n fois
?	0 ou 1 fois, optionalité d'un caractère

On peut spécifier le nombre de fois où un motif est répété, en mentionnant ce chiffre entre accolades :

- {k} : le motif est répété exactement k fois ;
- {k, n} : le motif est répété entre k et n fois ;
- {k,} : le motif est répété k fois ou plus

Ex. 19[0-9]{2}

EXPRESSIONS RÉGULIÈRES

\ sert à déspécialiser un métacaractère pour lui donner son sens littéral.

\. désigne un point

| disjonction, placé entre deux motifs, permet de décrire simultanément les segments de textes couverts par l'un ou par l'autre

L|le

Paris|Aix-en-Provence|Bordeaux

EXPRESSIONS RÉGULIÈRES

( ) permettent d'indiquer comment les éléments de l'expression doivent être groupés.

Fuchs, (C \.|Catherine) =

Exercices :
donner une ER

- n'importe quelle chaîne contenant *fred*

  ex : *Fred, frederick,* ou *Alfred*

- toute chaîne contenant au moins un *a* suivi d'un nombre quelconque de *b*

- toute chaîne contenant un nombre quelconque de barres obliques inverses suivies d'un nombre quelconque d'astérisques.

  ex : **, barney ***

- les lignes d'entrée qui mentionnent *wilma*. Faites en sorte qu'elles correspondent également à *Wilma*

- les lignes qui mentionnent à la fois *wilma* et *fred*

EXERCICES :
À EXPLIQUER LE SENS D'UN MOTIF

[^def]

[^n\-z]

a{5,15}

(fred){3,}

w{8}

^fred

rock$

a(bc)*

[]-]

^[^,]*

[\+?{}.]

EXPRESSIONS RÉGULIÈRES : SOUS CHAÎNE

\(\)

il est utile parfois d'identifier un certain type de chaîne pour pouvoir s'en servir dans la suite du traitement comme un sous programme.

c'est le principe des sous chaînes.

pour mémoriser une sous chaîne, on utilise la syntaxe \(ER\), cette sous chaîne sera identifiée par un chiffre compris par 1 et 9.

Exemple :

\([a-z] [a-z]*\) est une sous chaîne identifiant les lignes contenant une ou plusieurs lettres minuscules.

Pour faire appel à cette sous chaîne, on pourra utiliser \1.

TRAVAIL À FAIRE

- Objectif : annoter en entités nommées un article de Wikipédia en utilisant
  - les méthodes symboliques (expressions régulières, Unitex, CasEn)
  - Les méthodes par apprentissage (NLTK, SpyCy, Sem, etc.)
- Typologie :
  - Dates
  - Lieux
  - Personne
  - Institution
- Résultat :
  - Evaluer et analyser les résultats