Détection des Entités Nommées

N. Friburger
Master 2 Pro

Besoins

- Explosion de l'information
  - globalisation
  - quantité
- Accès à l’information
  - Diversité des documents, des supports, des langues
  - Recherche des documents pertinents ?

Besoins

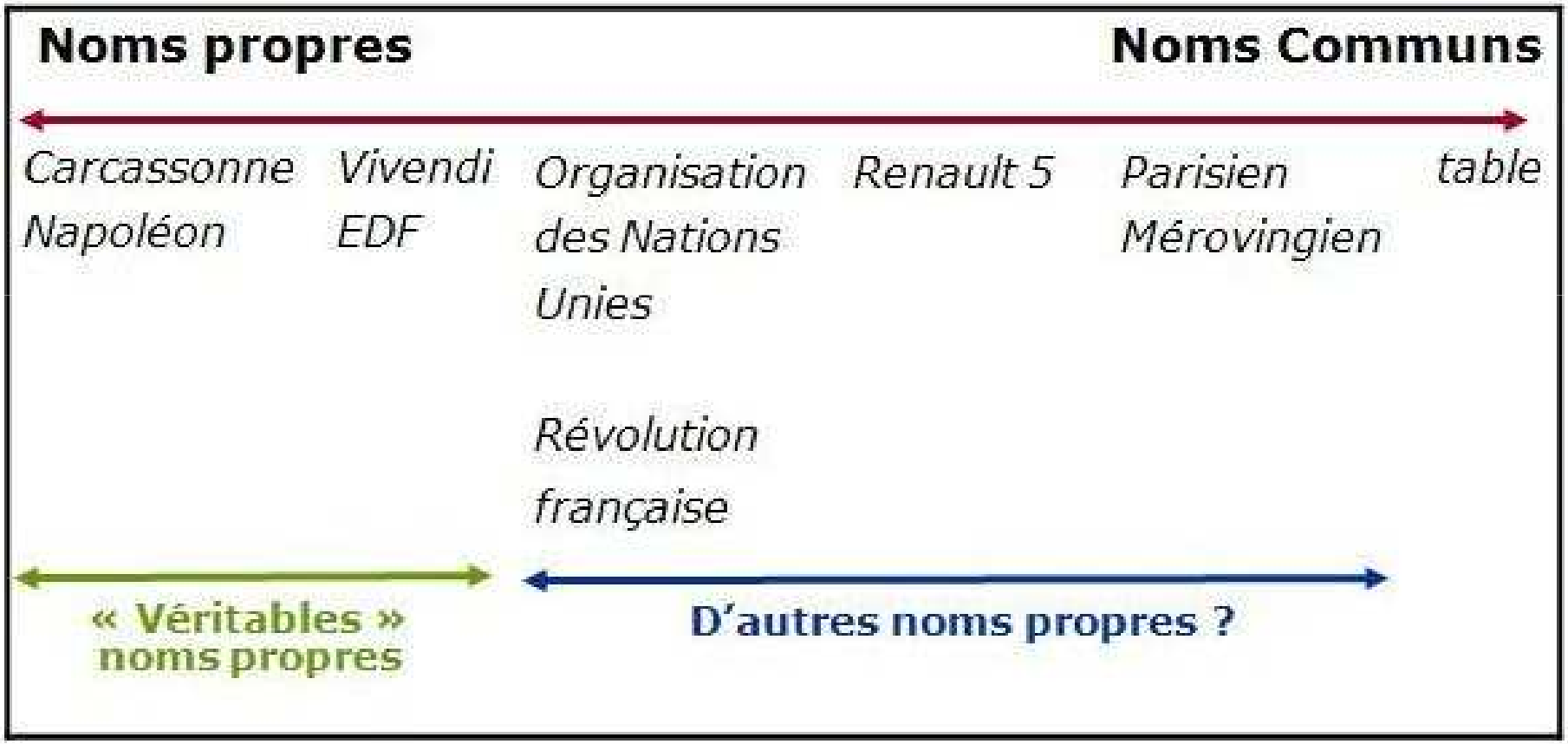
- Masse de textes à traiter
  - indexer l'information
    - pour la retrouver
  - marquer l'information dans les textes
    - pour aider à prendre des décisions
- Extraction d'information
  - tâches génériques
    - Reconnaissance des entités nommées
    - réponse aux questions
- Particularité des textes
  - information non structurées

Solutions

- **Accès à l'information**
  - Traitement automatique des langues, fouille de textes et accès multilingue
  - Recherche d'information
    - trouver les unités textuelles pertinentes
  - Classification et catégorisation de l'information
    - documents, termes
  - Extraction d'information
    - repérer les entités et les faits d'écrits dans un document
  - Résumé automatique
  - Extraction de terminologies multilingues (indexation)
  - Recherche d'information multilingue
  - Aide à la compréhension

Les entités nommées

- Qu'est-ce qu'un nom propre ? une entité nommée ?



Les entités nommées

- Unités lexicales particulières
  - noms de personnes, noms d'organisation, noms de lieux... dates,
  - unités monétaires, pourcentages...

- Reconnaissance des entités nommées
  - Identifier ces unités dans un texte
  - Les catégoriser

Détection des EN
Applications "internes"

- Pouvant servir à l'intérieur de applications "directes"

- Analyse syntaxique
  - Aide à la segmentation et à la morphosyntaxe
    - *Microsoft Corp.*
    - *M. Dupont*
    - *Airbus et Boeing sont en concurrence depuis une vingtaine d'années.*

- Aide à l'analyse syntaxique
  - *Il rencontrera cet émissaire envoyé par la <loc>Jordanie</loc> et l'<loc>Irak</loc>*
  - *Il rencontrera cet émissaire envoyé par la <loc>Jordanie</loc> et le <loc>partie irakien</loc>*

- Acquisition de dépendances "sémantiques"
  - *Ils se recontreront à <loc>Paris</loc>.*

Détection des EN
Applications "internes"

- Coréférence
  - <pers>Amélie</pers> *suit des cours d'informatique. Elle connaît bien le PHP.*

- Traduction
  - *<pers> Jack London</pers> is an american writer*

    Traduction :
    - **<pers> Jack Londres</pers> est un écrivain américain*
    - *<pers> Jack London </pers> est un écrivain américain*

- Désambiguïsation lexicale
  - *<pers> Paris Hilton</pers> est une starlette peroxydée américaine*
  - *La chaîne d'hotel <org>Hilton</org> est très luxueuse*
  - *<loc>Paris</loc> est une très belle ville*

Détection des EN
Applications "directes"

- L'extraction d'information et la veille
  - Remplir des bases de données avec des entités / types d'entités donnés
- Signaler de nouveaux documents concernant cette entité ou ce type d'entités
- Tâche de questions/réponses
  - Permet d'identifier le type de réponse attendu
- L'anonymisation

Les entités nommées

- *Texte*

Lundi matin, le président du Parti républicain, M. Gérard Longuet, a déclaré, sur RTL, que " le FLN a complètement raté sa mission ", mais que " le FIS n'est pas un élément inéluctable en Algérie ".

- *Après reconnaissance et catégorisation*

Lundi matin, le président du Parti républicain, <pers>M. Gérard Longuet</pers>, a déclaré, sur <org>RTL</org>, que " le <org> FLN </org> a complètement raté sa mission ", mais que " le <org> FIS </org> n'est pas un élément inéluctable en <loc> Algérie</loc> ".

Les entités nommées

- *Après reconnaissance et catégorisation / résultats plus fin*

  Lundi matin, <pers><fonction>le président du Parti républicain,</fonction> M. Gérard Longuet</pers>, a déclaré, sur <org>RTL</org>, que " le <org> FLN </org> a complètement raté sa mission ", mais que " le <org> FIS </org> n'est pas un élément inéluctable en <loc> Algérie</loc> ".

Les entités nommées

- Un autre exemple

Une lettre du <NE type="pers.hum"><NE type="fonc.pol">préfet de Corse</NE> </NE> A la suite de notre article " Polémique en <NE type="loc.admi"> Corse</NE> autour de la base marine du futur parc marin des bouches de <NE type="loc.admi">Bonifacio</NE> " (<NE type="prod.doc">Le Monde</NE> du <NE type="time.date">4 août</NE>), <NE type="pers.hum"><NE type="pers.hum.tit">M.</NE> Bernard Bonnet</NE>,<NE type="fonc.pol"> préfet de Corse</NE>, tient à apporter les précisions suivantes : 1) Je n'ai pas saisi le procureur de la République d'une plainte, mais d'une demande de démolition d'une construction édifiée en complète illégalité dans une zone inconstructible. <NE type="loc.geo">Car</NE> comme vous le savez, seule l'autorité judiciaire peut ordonner la démolition d'une construction. 2) Le procureur de la République n'a aucune compétence pour annuler un acte administratif, comme l'écrit votre correspondant. La base nautique a été édifiée sur le fondement d'une " autorisation de travaux " déclarée par le <NE type="fonc.pol">maire de Bonifacio</NE> qui n'avait pas qualité pour le faire, les permis de construire sollicités par des établissements publics étant délivrés uniquement par le <NE type="fonc.pol">préfet</NE>. L'imbroglio juridique décrit par <NE type="pers.hum"><NE type="pers.hum.tit">M.</NE> Silvani</NE> n'existe donc pas. 3) " Le petit bâtiment " évoqué par votre correspondant représente une superficie de plus de <NE type="amount.phy.area">400 m²</NE> . Le bâtiment préexistant détruit par l'explosif <NE type="time">en 1991</NE> est évidemment postérieur à la promulgation de la loi littoral, qui date de 1976. 4) S'il est vrai que la décision de créer un parc marin a été prise par un comité de pilotage présidé par <NE type="pers.hum">Claude Erignac</NE>, il n'a jamais été acté qu'une telle prise en compte dispensait de respecter les lois et réglements régissant la délivrance des actes d'urbanisme et les marchés publics. Mon prédécesseur avait d'ailleurs été contraint de déférer devant le tribunal administratif de <NE type="loc.admi">Bastia</NE> le marché de construction de la base.

Les problèmes à résoudre

- *Le choix des classes*
  - Les classes selon MUC
    - Les classes "ENAMEX" définies pour MUC:
      - ORGANISATION
      - LIEU
      - PERSONNE
    - TIMEX (date, expressions temporelles)
    - NUMEX (valeur monétaire, pourcentage...)

- En français
  - les classes Ester
    - <NE type = "pers.hum" > … </NE>
  - Les classes prolex
    - <anthroponyme …>

Les problèmes à résoudre

- *La portée des classes*
  - L'exemple des personnes
    - *Lionel Jospin* *les Kennedys* *la famille Kennedy*
    - *l'épouse Chirac*
    - *les Windsor* *les frères Coen*
    - *Zizou*
    - *les Démocrates* *les italiens* *les Talibans* *les Peuls*
    - *Bison futé* *Mickey* *Zorro* *Hercule*
    - *le Prince Charmant* *l'ours Colargol* *Milou*
    - *St. Nicolas* *Vichnu*
    - ?

Les problèmes à résoudre

- *La coordination*
  - Combinaisons de syntagmes
    - *Jacques et Bernadette Chirac vont au Japon la semaine prochaine (ellipse partielle)*
    - *M. et Mme. Chirac étaient au Japon. (ellipse totale)*
    - *Les banques centrales européenne et américaine sont intervenues ...*
  - Une ou plusieurs entités ?
    - *<pers>Jacques</pers> et <pers>Bernadette Chirac</pers> vont au Japon la semaine prochaine.*
    - <pers type="collectif"> *Jacques et Bernadette Chirac </pers> vont au Japon la semaine prochaine.*
    - <pers>*Jacques Chirac</pers> et <pers>Bernadette Chirac </pers> vont au Japon la semaine prochaine.*

Les problèmes à résoudre

- *L'imbrication*
  - Combinaisons de syntagmes :
    - *L'Université de Tours.*
    - *Le conseil de la vie universitaire de l'université de Tours*
    - *Microsoft Corp.*
  - Une ou plusieurs entités ?
    - <org>*L'Université de Tours*</org>
    - <org>*L'Université de <loc>Tours</loc>*</org>

Les problèmes à résoudre

- *Les frontières des EN*

Le Palais Bourbon	*Les Rolling Stones*
Le téléphone sonne	
La Mecque	
le cardiologue Dupont	
Monsieur Fillon	*Professeur Paolucci*
la candidate Ségolène Royal	*le président Nicolas Sarkozy*
l'Abbé Pierre	*Benoît XVI*
George W. Bush Jr.	*Lord Liverpool*
la secrétaire d'état Rama Yade	
Sir Paul Mc Cartney	*les célèbres Beatles*
?	

Découpage en phrases

- Découpage en phrases
  - Ambiguïté des points et des majuscules
    - Il a licencié F. Durant pour cette faute.
    - L'action Vivendi coûte 20 F. Durant le dernier trimestre, elle a diminué de 50%.

Les problèmes à résoudre

- *Les variantes*
  - *Jacques Chirac*
    - *Président Jacques Chirac*
    - *Chichi*
    - *Chirac*
    - *l'ancien président français*
  - *Elisabeth II*
    - *la reine d'Angleterre*
  - *l'Association Sportive de Saint-Etienne*
    - *l'ASSE*
    - *le club forézien*
  - *Le Stade Toulousain*
    - *Toulouse (métonymie)*

Les problèmes à résoudre

- *La polysémie*
  - *Orange*
    - la ville ?
    - la société ?
  - *Vienne*
    - la ville en France
    - la ville en Autriche
  - *Leclerc*
    - le Maréchal
    - l'homme d'affaire
    - le char
    - le supermarché
  - *Jacques Chirac*
    - le président de la République
    - le maire de Paris
    - le groupe financier

Détection des EN

- La majuscule ?
  - Metro Goldwyn Mayer
  - Albert Einstein

- Indice insuffisant, pourquoi ?
  - Le premier mot d'une phrase porte la majuscule
- Problème de la limite droite
  - Organisation des Nations Unies

Détection des EN

- Le problème de la limite droite
  - Les solutions proposées
    1. Chercher une séquence contiguë de mots capitalisés
       - (Arrêt au premier mot non capitalisé ou à une virgule)
    2. Grammaire syntaxique de l'extension à droite des noms propres
       - ex : Institut national de la recherche agronomique
       - mais ex : Organisation mondiale de la santé performante
    3. Utiliser la syntaxe et le lexique
       - grammaire de l'extension des noms à droite utilisée avec un dictionnaire d'adjectifs et de noms pouvant se trouver dans les noms d'organisations

Détection des EN

- Le lexique
  - Sert à le reconnaître mais aussi à le catégoriser à l'aide de preuves (McDonald, 1996)
    - Preuve interne
    - Preuve externe

Détection des EN

- Les preuves internes
  - Majuscule (à manipuler avec précaution)
  - Pour les personnes : les prénoms, abréviations de prénoms, chiffres romains
    - **André** Malraux
    - ***A.*** *Jospin*
    - *Jean-Paul* ***II***
  - Pour les lieux et organisation : mots classifiant
    - *la* ***Société*** *Générale*
    - *Microsoft* ***Inc.***
    - *le Mont* ***Blanc***
    - *la* ***mer*** *rouge*
    - Wall Street **Journal**
    - **Organisation** des Nations Unies
  - Pour les organisations : sigles, esperluettes
    - *Crédit Agricole* ***SA***
    - *C**&**A*

Détection des EN

- Les preuves internes
  - Description de la structure des prénoms
    - Ex : Jean, Jean-Pierre, J.-P., George W., etc.
  - Structure des patronymes
    - Patronymes simples
      - Ex : Dupont
    - Patronymes composés avec particule étrangère
      - Ex : Mac Donnell-Douglas, O'Ryan, von Bulow, Da Silva, etc.
    - Patronymes français à particules
      - Ex : Dupont de Nemours, de la Fontaine.
  - Coordinations de noms de personnes
    - Ex : L'assemblée générale de la MIRCEB est présidée par MM. Foucaud, Barriolade, Fromaget, Delaunay, Gosselin, Vincent,

Détection des EN

- Les preuves externes
  - Contexte d'apparition des entités nommées
    - Informations supplémentaires ou propriétés spécifiques
  - Pour les personnes (titre, grade, métier etc.) :
    - **Monsieur** Jospin
    - **Mme** Denise
    - **Général** Leclerc
    - l'**entraîneur** Aimé Jacquet
    - Le **juge** van Ruymbeke
  - Pour les organisations
    - le **groupe** Carrefour
    - la **filiale de** Vivendi
  - Pour les lieux
    - le **fleuve** amazone
    - la **ville de** Tours
    - la **planète** Mars

Détection des EN

- Problème de conflit preuve interne/preuve externe
  - Ex : La société Hugues Aircraft
  - Preuve interne insuffisante
    - Solution : Priorité à la preuve externe !

Détection des EN

- Utilisation de dictionnaires
  - prénoms, marques, sigles, lieux …
  - MAIS lorsqu'on utilise des dictionnaires
    - s'ils sont trop petits, ils ne servent à rien !
    - s'ils sont trop gros, ils introduisent trop d'ambiguités

  - Détection des noms de lieux
    - Preuves internes et externes
    - mais surtout, de part leur "constance" et leur relativement faible quantité,
      - on utilise des dictionnaires les contenant !!!

Systèmes de détection des EN

- Systèmes à base de règles
  - Utilisation de preuves internes et externes
  - Dictionnaires
  - Description de règles : expressions régulières
    - Lisibilité
    - Évolutivité (ajout de vocabulaire, de règles…)
    - Pas de corpus d'apprentissage
  - Investissement en terme de description
  - Privilégiés pour le travail sur l'écrit
  - Résultat Rappel et précision > 90%
  - L'existant
    - Funes, PNF, LaSIE, Nominator, Exoseme, ThingFinder, Facile, etc.
    - Sur le français : Exoseme (Wolinski et al., 1995), Thingfinder (Trouilleux, 1997),CASSEN (Friburger,2002)

Systèmes de détection des EN

- Systèmes à apprentissage
  - Résultat : des règles logiques, un arbre de décision, un modèle numérique...
    - Nécessitent de larges corpus annotés
    - Pas toujours possible d'intervenir sur les résultats après coup
  - Minimisent le travail de description
    - Meilleure robustesse sur des données bruitées
    - Mieux adaptés aux formats de données utilisées pour l'oral (treillis de mots, modèle de langages…)
    - Notion de corpus d'apprentissage mieux admise
  - Privilégiés pour l'analyse de l'oral
  - Résultats très variables de 50% à 90% de rappel
  - L'existant
    - Alembic, BBN IdentiFinder, MENE, Answer extraction, etc.
    - Sur le français : (Béchet et al., 2000)

Systèmes de détection des EN

- Approches mixtes
  - Apprentissage de règles puis révision par un expert
  - Élaboration de règles par un expert puis extension automatique de la couverture
  - Avantages et inconvénients des 2 précédents systèmes