

Le Traitement des Entités Nommées

DEFINITION, RESOURCES, MÉTHODES, APPLICATIONS

Maud Ehrmann ¹ **Sophie Rosset** ²

11 Juillet 2018

Session 1: Cours (1h30)

Session 2: Cours (1h45)

Session 3: TP, Elaborer un système de reconnaissance d'EN (1h30)

Session 4: TP, Evaluer un système de reconnaissance d'EN (1h30)

¹EPFL-DHLAB, Lausanne, Switzerland

²LIMSI-ILES, Orsay, France

Plan du cours

Estimations (pour nous, ne sera pas affiché)

- **Session 1:** 1h30min (pauses comprises)
 1. Contexte et Applications (25min)
 2. Définition (15min)
 3. Pause (15min)
 4. Resources (35min)
- **Session 2:** 1h30 (pauses comprises)
 1. Reconnaissance et classification (35min)
 2. Pause (10min)
 3. Liaison (20min)
 4. Evaluation (20min)

Plan du cours

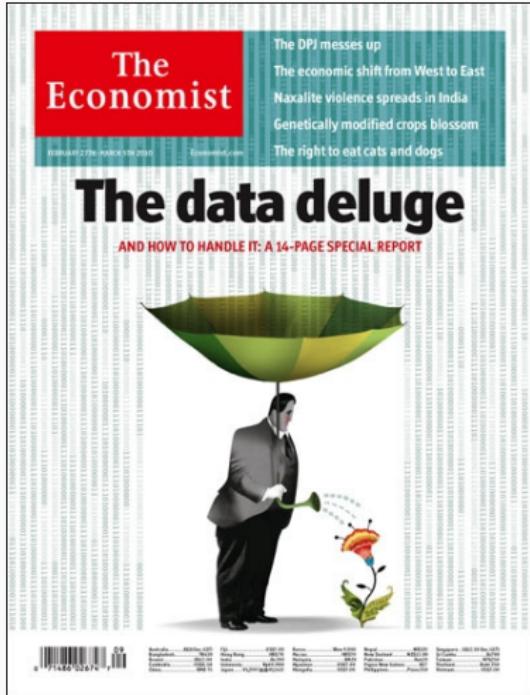
1. Contexte et Applications
2. Définition
3. Resources
4. Reconnaissance et classification
5. Liaison
6. Evaluation

1. Contexte et Applications

1. Contexte et Applications

1.1 Introduction

Contexte



Données

Quoi: TOUT

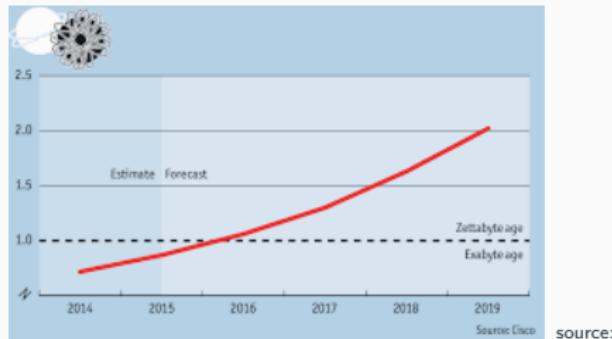
i.e. texte, images, audio publiés via sites de news, médias sociaux, plateformes collaboratives, smartphones, capteurs, etc.



Profil (1/2)

Combien: croissance astronomique

- quantité: double tous les 2 ans
- trafic: entré dans l'ère des zettabyte en 2016 (1 trillion gigabytes)
- stockage: prévision de 44 zettabytes en 2020



source:

<http://www.theworldin.com/article/12107/charting-change>

Profil (2/2)

Nature: 80 à 90% des données sont **non structurées**, i.e. sans modèle ni format pré-définis. **Défis**

- stockage de plus en plus coûteux
- mais surtout: exploiter les données, **extraire l'information utile**

Mise au point

Donnée	Information	Connaissance
description élémentaire d'une réalité	données avec un sens construisant une représentation de la réalité	informations avec une vérité
<i>mesure des températures dans une station météo</i>	<i>une courbe donnant l'évolution des minima et maxima moyens en un lieu donné, par mois</i>	<i>le fait que la température sur terre augmente du fait de l'activité humaine</i>
série d'articles journalistiques	<i>nom de personnes et leurs polarités</i>	<i>opinion des médias vis-à-vis de personnalités</i>

inspiré de: http://www.college-de-france.fr/site/serge-abiteboul/_inaugural-lecture.htm

Données semi-structurée

Cannes Film Festival

From Wikipedia, the free encyclopedia

Coordinates: 43°33'03.10"N 7°01'02.10"E

The Cannes Festival (*/kænɪʃ/*) (French: *Festival de Cannes*), named until 2002 as the International Film Festival (*Festival international du film*) and known in English as the Cannes Film Festival, is an annual film festival held in Cannes, France, which previews new films of all genres, including documentaries, from all around the world. Founded in 1946, the invitation-only festival is held annually (usually in May) at the Palais des Festivals et des Congrès.^{[1][2][3]}

On 1 July 2014, co-founder and former head of French pay-TV operator Canal+ Pierre Lescure took over as President of the festival. The Board of Directors also appointed Gilles Jacob as Honorary President of the festival.^{[4][5][6]}

The 2016 Cannes Film Festival took place between 11 and 22 May 2016. Australian film director George Miller was the President of the Jury. *I, Daniel Blake*, directed by British director Ken Loach, won the Palme d'Or.

In 2017, The Festival de Cannes will celebrate its 70th anniversary edition from May 17 to 28.

Contents [hide]

- 1 History
- 2 Impact
- 3 Programmes
- 4 Juries
- 5 Awards
- 6 See also
- 7 References
- 8 Further reading
- 9 External links



Festival de Cannes

@Festival_Cannes



Follow

In French theaters today, testimonies from Ugandan ex-child soldiers : Wrong Elements by Jonathan Littell #SpecialScreening in #Cannes2016

Cannes Film Festival



FESTIVAL DE CANNES



Location Cannes, France
Founded September 20, 1946
Awards Palme d'Or, Grand Prix
Website festival-cannes.com

*mais la plupart du temps,
l'information est ‘cachée’ dans les textes*

Données non-structurées

“On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar.”

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th Festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar.

Monica Belucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present *Under Suspicion* by Stephen Hopkins. She returned two years later with Gaspar Noé's steamy *Inferno*, which entranced the Croisette with its unforgettable scenes.

Monica Belucci was a member of the Jury in 2006 under the presidency of Wong Kar-wai. In the following years, Belucci returned to Cannes for the Official Selection with Marco Tullio Giordana's *What Blood*, and *Don't Look Back* by María de Varn. In 2014, she was back on the Croisette to present *The Wonders* by Italian director Alice Rohrwacher, which picked up the Jury Grand Prix.

Belucci's film career demonstrates her ease across a range of genres with outstanding performances in both comedy and drama, based on eclectic and daring artistic choices. She has films for a number of prestigious directors including Bertrand Blier, Danièle

source: www.festival-cannes.com

Information ‘cachée’ dans les textes

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar.

PERSON, ORGANIZATION, TIME-EXPR, EVENT

Extraction d'Information (EI)

L'objectif est d'extraire des informations structurées à partir de textes non structurés, c-a-d:

- identifier et catégoriser des fragments d'information
- les relier avec des bases de connaissances
- les aggréger pour extraire d'autres informations

Principales tâches en EI

- **traitement des entités nommées:**

reconnaissance, catégorisation et désambiguisation

- *Monica Belluci* et *Pedro Almodovar* sont des PERSON.
 - *Monica Belluci* $\xrightarrow{\text{reference}}$ http://dbpedia.org/page/Monica_Bellucci

- **traitement des expressions temporelle:**

extraction et normalisation

- *from 17 to 28 May 2017* est une DURATION
 - *from 17 to 28 May 2017* → [17-05-2017, 28-05-2017]

- **extraction d'événements**

- *70th Festival de Cannes* est un FACTUAL, RECURRING EVENT
 - *70th Festival de Cannes* $\xrightarrow{\text{instance_of}}$
https://en.wikipedia.org/wiki/Cannes_Film_Festival

- **extraction de relations:**

- *70th Festival de Cannes, tookPlace, [17-05-2017, 28-05-2017]*

1. Contexte et Applications

1.2 Un peu d'histoire

De la compréhension à l'extraction

- **1980s:** objectif **compréhension automatique** de textes
- Un projet **trop ambitieux** face à des difficultés techniques et théoriques:
 - faible couverture des grammaires
 - trop d'ambiguïtés non résolues
 - difficultés à collecter, représenter et manipuler les connaissances

→ approche générique de la compréhension de textes est encore une **utopie**
- **1990s:** **décomposition de la tâche** de compréhension
 - se focalise sur des éléments précis d'intérêt
 - un modèle est défini à l'avance en fonction de l'application
 - analyse locale (10-20% du texte nécessaire).

La série des conférences MUC

- *Message Understanding Conference*
- Cycle de 7 campagnes d'évaluation entre 1987 et 1998
- Initié par la Division pour la Recherche et le Développement de la Marine américaine
- Financé par le DARPA (Defense Advanced Research Project Agency)

Evolution des conférences MUC

Phase 1: cycle exploratoire

- **1987 (MUC-1)** pas de tâche précise, rapports militaires sur des opérations navales en style télégraphique;
- **1989 (MUC-2)** définition de formulaires prédéfinis (*templates*) devant être complétés (**10** champs); définition de mesures d'évaluation (precision et rappel).

Evolution des conférences MUC

Phase 2: remplissage de formulaires de plus en plus complexes

- **1991 (MUC-3)** dépêches de presse sur événements terroristes en Amérique centrale et du sud; formulaire avec **18** champs.
- **1992 (MUC-4)** idem, **24** champs
- **1993 (MUC-5)** tâches plus complexes, test sur domaines nouveaux, 2 langues, 11 formulaires **48** champs hiérarchisés

MUC 3: Définition de la tâche de compréhension

Etant donné un document, il était demandé de:

- repérer des événements,
- repérer les éléments s'y rattachant,
- "normaliser" ces éléments,
- remplir un formulaire descriptif.

e.g. pour chaque événement, il fallait trouver son type, sa date, les agents impliqués, le lieu etc.

Formulaire MUC-3

19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador : San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

todo: ajouter ref

MUC 5: extension de la tâche de compréhension

- plus de domaines (2), de langues (2), de champs dans le formulaire (48!)
- formulaires différents selon le domaine
- structuration des éléments composants le formulaire [GS95]
 - un objet par événement
 - chaque objet peut pointer sur une liste d'objet
 - un objet représente un participant à un événement

Evolution des conférences MUC

Phase 3: reformulation des objectifs et définition de sous-tâches

- **1995 (MUC-6)**

- technologies et composants indépendants
 - définition de la sous-tâche dédiée aux "entités nommées"
 - systèmes portables
 - définition de template génériques
 - considération des "briques de base de la compréhension"
 - co-reference, désambiguisation lexicale, structure argument-prédicat

- **1997 (MUC-7)** continuation

MUC 6: point de départ des travaux sur les EN

- Définition de la notion d'entité nommée
- Définition de la tâche de reconnaissance des EN

Ensuite: MET, IREX, CONLL, ACE, ESTER, HAREM, EVALITA,
GERMEVAL, TREC, TAC, etc.

1. Contexte et Applications

1.3 Définition courante

Entités nommées: première définition (TAL)

- des éléments "d'intérêt", généralement de type *Personne*, *Organisation*, *Lieu*
- des unités référentielles qui sous-tendent la sémantique des textes.

Entités nommées: différentes tâches

1. **reconnaissance**: détecter, repérer des entités nommées dans les flux textuels (on pose les frontières dans le texte)
2. **classification**: catégoriser les éléments reconnus selon des catégories sémantiques pré-définies (on affecte un type)
3. **désambiguïsation/liaison**: lier les mentions d'entités à une référence unique (on lie à une référence)
4. **extraction de relation**: découvrir des relations entre entités (*father-of, born-in, alma mater*)

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar. [...] Monica Bellucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present *Under Suspicion* by Stephen Hopkins.

Application de Stanford NER

On the invitation of the **Festival de Cannes**, the Italian actress **Monica Belucci** has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de **Cannes** to be held from 17 to 28 **May 2017**, under the presidency of Spanish filmmaker **Pedro Almodovar**. [...] **Monica Bellucci**'s friendship with the **Festival de Cannes** goes back a long way: in **2000**, she walked up the steps for the first time to present *Under Suspicion* by **Stephen Hopkins**.

PERSON, ORGANIZATION, LOCATION, DATE

Plus d'information?

On the invitation of the **Festival de Cannes**, the Italian actress **Monica Belucci** has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de **Cannes** to be held from 17 to **28 May 2017**, under the presidency of Spanish filmmaker **Pedro Almodovar**. [...] **Monica Bellucci**'s friendship with the **Festival de Cannes** goes back a long way: in **2000**, she walked up the steps for the first time to present **Under Suspicion** by **Stephen Hopkins**.

Désambiguisation et relations

On the invitation of the **Festival de Cannes**, the Italian actress **Monica Bellucci** has agreed to play the role of Mistress of the Ceremonies of the 70th festival de **Cannes** to **May 2017**, under the presidency of Spanish **Carles Puigdemont**. [...] **Monica Bellucci**'s friendship with **Stephen Hopkins** goes back a long way: in **2000**, she walked up the red carpet to present *Under Suspicion* by **Stephen Hopkins**.



DBpedia

About: [Monica Bellucci](#)

An Entity of Type : person, from Named Graph : [http://dbpedia.org](#), within Data Space : [dbpedia.org](#)

WIKIDATA

Item Discussion

Monica Bellucci (Q81819)

Italian actress

1. Contexte et Applications

1.4 Applications

- Étiquetage morpho-syntaxique et analyse syntaxique de surface
 - HyOx, Inc.
 - Seat and Porsche has fewer registrations in July 1996.
- Analyse syntaxique
 - *He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to Egypt and Jordan.*
 - *He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to Egypt and Likud party.*

Applications ‘internes’ au TAL (2/4)

- **Analyse en dépendances**

*They met in **Bagdad**.* → LOCATION(*met*, *Bagdad*)

- **Coreference**

*John bought a new computer. **It** was able to train the model.*

- **Traduction**

Jack London was an american writer. London is a busy city.

- **Désambiguisation lexicale**

- *It is difficult to leave Paris on Friday evenings.*
 - *Some wonder if they will leave the Socialist party.*

Quelle est la signification de 'leave' ?

- **Désambiguïsation lexicale**

It is difficult to leave Paris on Friday evenings.

→ leave = "go away from a place" (#1 WordNet)

Some wonder if they will leave the Socialist party.

→ leave = "remove oneself from an association with or participation in" (#8 WordNet)

Applications

- **Extraction d'information et 'media monitoring'**
 - population de bases de connaissances avec des informations relatives à des entités
 - alertes sur certains sujets ou entités
- **Clustering de documents cross-lingue**

Les documents mentionnant les mêmes entités ont de fortes chances d'être reliés.
- **Résumé automatique**

Les EN sont des 'ancres' informationnelles aidant à identifier les éléments clés d'un texte

contexte: à retenir

- La notion d'EN est apparue dans les **années 90** lors de campagnes d'évaluations sur la **compréhension de documents**
- Les EN ont rapidement pris une place importante et sont devenues un **pivot central** pour les systèmes d'analyse automatique des textes.

2. Définition

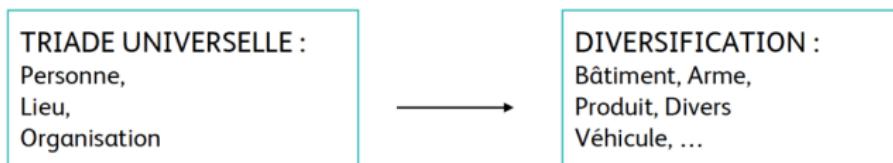
A quoi correspondent vraiment les entités nommées? Comment les définir?

2. Définition

2.1 Difficultés d'appréhension

Les EN dans le monde : le problème de la catégorisation

- Le choix des catégories



- La détermination de ce qu'elles recouvrent

Catégorie PERSONNE :

Lionel Jospin	les Démocrates	Bison Futé
les Windsors	les Talibans	le Prince Charmant
la famille Kennedy	Zorro	l'épouse Chirac
les frères Cohen	St Nicolas	...

→ catégorisation instable

Les EN dans le texte : le problème de l'annotation

- **Combinaisons de syntagmes : une ou plusieurs entités ?**

Les Banques centrales américaine et européenne ont décidé...

Bill et Hillary Clinton

l'Université de Corte

- **Un syntagme : quelles frontières ?**

la candidate Ségolène Royal, Professeur Paolucci

George W. Bush Jr., La Mecque, l'Abbé Pierre

- **Une entité : quelle unité lexicale ?**

Jacques Chirac, Monsieur Chirac, le Président Jacques Chirac,

le Président français, le Président de la République française, Chichi

→ **caractérisation imprécise, diversité des mentions**

Les EN dans la langue : le problème des « polysémies »

- **Homonymie**

Orange a invité M. Hollande.

- **Métonymie**

Leclerc a fermé ses magasins en Rhône-Alpes.

- **“Facettes”**

Le candidat Sarkozy, devenu chef de l'Etat, a changé de position sur la présence française au sein de la force internationale.

→ **polyréférentialité**

- **Hétérogénéité des réalisations**

Les entités nommées ne se limitent pas à une catégorisation, une mention, une interprétation.

- **Hétérogénéité des points de vue**

- Formules définitoires sous la forme d'énumérations
- Caractérisations diverses (sens, forme)

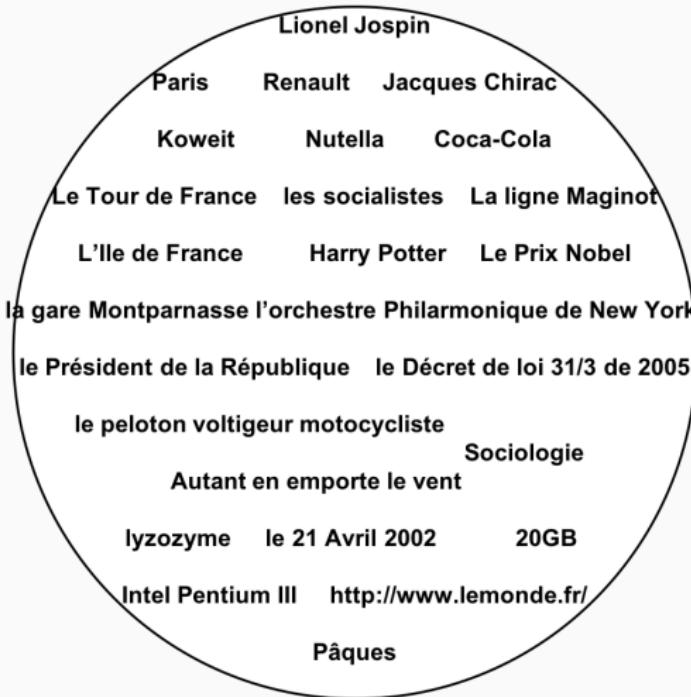
→ **Question** : Que sont les entités nommées ?

2. Définition

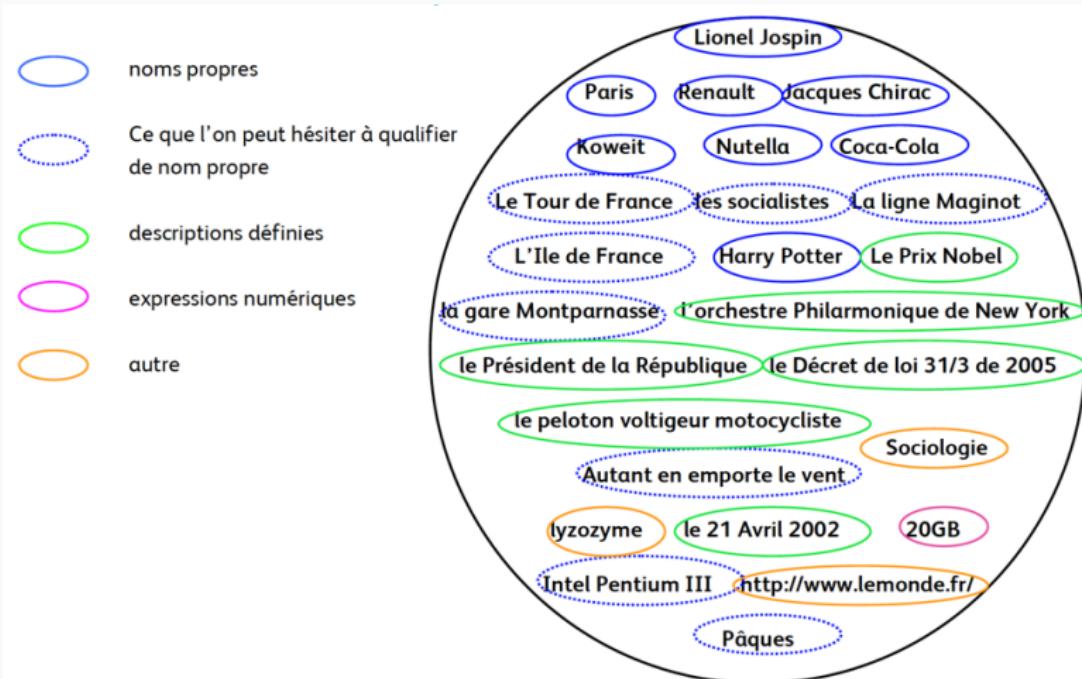
2.2 Vers une définition des entités nommées

Le “matériau” de départ

Unités lexicales
considérées
comme des entités
nommées



Le “matériau” de départ



Proposition de définition

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

Questions que l'on s'est posées :

- Comment définir un objet TAL ?
- Que sont les noms propres et les descriptions définies ?
- Que devient le cadre linguistique du sens et de la référence en TAL ?

Considération des aspects linguistiques

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute **expression linguistique qui réfère à une entité unique** du modèle de manière **autonome** dans le corpus.

Sens et référence en linguistique

- La **référence** désigne le lien qui existe entre une expression linguistique et l'élément du réel auquel elle renvoie.
- Le **sens** détermine les caractéristiques qu'une entité doit satisfaire pour pouvoir être désignée par telle ou telle expression.
- Un **modèle hétérogène** du sens (G. Kleiber)
 - sens descriptif (*oursin, table*)
 - sens instructionnel (*je*)
- **Comprendre** grâce à :
 - des connaissances lexicales
 - des connaissances sur le contexte
 - des connaissances sur le monde

- **Le nom propre réfère à un particulier**
 - **nomination d'un particulier** (Felix) vs. nomination d'une classe conceptuelle (chat)
 - **unicité** : une individualité considérée comme unique au sein d'une catégorie d'exsistants
 - **unité** : une individualité considérée comme formant un tout reconnaissable
- **Les descriptions définies**
 - présupposition d'existence et d'unicité
le président de la République, le père de Charles II, le marronnier
Une description de la forme "le tel et tel" présuppose qu'il existe une et une seule entité qui soit telle et telle

Comment s'opère la référence à une entité unique ?

Noms propres

- sens instructionnel dénominatif → connaissance d'une convention
- dénomination non contingente → désignateur rigide
- dénomination plus ou moins descriptive (*Massif Central*)

Descriptions définies

- sens descriptif
- descriptions définies complètes et incomplètes
le président, le président de la République française en 2003

- **L'ensemble 'entités nommées' n'est pas réductible à une catégorie linguistique**

'Plus que les noms propres et moins que les descriptions définies'

- **Caractérisation d'un comportement référentiel**

Référence à une entité unique et autonomie référentielle

Jacques Chirac, le Président de la République, le costume bleu du président

→ La perspective linguistique ne suffit pas

Considération des aspects liés au TAL

Etant donné un **modèle applicatif** et un **corpus**, on appelle entité nommée toute **expression linguistique qui réfère à une entité unique** du modèle de manière **autonome dans le corpus**.

- **Caractérisation de la référence en TAL**

- restriction
- représentation

La référence en TAL désigne le lien qui existe entre une expression linguistique et l'élément du modèle auquel elle renvoie.

- **Comprendre en TAL grâce à**

- des ressources lexicales
- des ressources encyclopédiques
- des informations sur le contexte (issues du corpus)

- **Articulation sens–référence en TAL**

- entre le langage et le modèle
- trois mécanismes : segmentation, classification et reformulation

Illustration

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

		le président de la République en 2005
Laguna	Jacques Chirac	Napoléon III
le président	je	30°
	l'Empereur des Français	2028hPa
Ivan	le président de la République en 2007	l'été 2004
	l'ouragan	Louise Colet

Application : générique « typique »

Modèle : Personnes, Lieux, Organisations

Corpus : journalistique français de 1998 à 2008

Illustration

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

Laguna	le président de la République en 2005		
	Jacques Chirac	Napoléon III	
le président	je		30°
	l'Empereur des Français		2028hPa
Ivan	le président de la République en 2007		
	l'ouragan	Louise Colet	l'été 2004

Application : étude sur le climat

Modèle : températures, mesures atmosphérique, ouragan, dates, périodes, ...

Corpus : totalité des observations météorologiques sur une période données

Illustration

Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.

		le président de la République en 2005
Laguna	Jacques Chirac	Napoléon III
	le président	je
		30°
	l'Empereur des Français	2028hPa
Ivan		le président de la République en 2007
	l'ouragan	Louise Colet
		l'été 2004

Application : « littéraire »

Modèle : personnes, lieux, événements

Corpus : correspondance de Flaubert

De la linguistique au TAL, spécification d'un cadre théorique pour les EN :

- perspective linguistique : non réductibles à une catégorie mais caractérisables par un comportement référentiel
- perspective TAL : existent relativement à un modèle applicatif précis

→ Pas d'entité nommée « en soi », seulement des critères linguistiques et un modèle.

Conséquences

- point de vue général : explication de l'hétérogénéité et de la variabilité de l'ensemble 'entités nommées'
- point de vue pratique : critères de décision pour annoter
- point de vue méthodologique : besoin impératif d'expliciter le modèle



(15min break)

3. Resources

De quoi a-t-on besoin pour traiter les entités nommées?

1. **Typologies**, pour définir d'un cadre sémantique
2. **Corpus annotés**, pour servir de référence (évaluation) et d'illustration
3. **Lexiques et bases de connaissances**, pour donner des informations sur les éléments à traiter (entraînement)

3. Resources

3.1 Typologies

Typologies: une façon de structurer

- Une typologie (ou *tagset*) est une **formalisation descriptive** des catégories d'EN à prendre en compte:
 - quoi reconnaître (cibler des éléments appartenant à des catégories spécifiques)
 - comment le représenter (pour un élément, choisir une catégorie parmi d'autres)
- De **multiples variations** en fonction des domaines et des applications
 - différences de catégories
 - différences de structure
 - différences sur la définition de ce que recouvrent les catégories

Catégorie : une notion centrale

Définitions:

- *une classe dans laquelle on range des objets de même nature* (Petit Robert)
- *un ensemble de choses qui ont un certain nombre de caractères en commun* (TLFi)

Pour les EN:

- les traits communs entre les objets sont de nature sémantique.
- la détermination des catégories revient donc à spécifier des classes sémantiques.
- concrètement, les catégories sont les étiquettes utilisées pour annoter les EN, c'est à dire leur *type*

Comment déterminer des catégories?

Approches:

- *top-down*: on a une idée, on définit, et voilà.
- *bottom-up*: les catégories émergent des données
- mixte: on a des idées, on confronte au données, on remanie.
- utilisation de ressources : catégories issues des infobox de Wikipedia, de son équivalent sémantique DBpedia (200 classes), plus récemment de Wikidata.

Dans les faits:

- très peu d'explications données sur l'élaboration des typologies
- influence thématique des financeurs
- seul Sekine [SSN02] a détaillé sa méthodologie de définition de sa typologie (200 !).

Typologie MUC

- **noms propres** (ENAMEX) : lieux, personnes, organisations,
- **expressions numériques** (NUMEX) : dates et heures (expressions absolues), montants monétaires et pourcentages.

Types	Exemple	Contre-exemple
ORG	DARPA	our university
PERS	Harry Schearer	St. Michael
LOC	U.S.	53140 Gatchell Road
MONEY	19 dollars	ça en coûte 19
TIME	8 heures	la nuit dernière (+ MUC7)
DATE	en juillet	en juillet dernier (+ MUC7)

<date> 19 March </date> - A bomb went off this morning near a power tower in <loc> San Salvador </loc> leaving a large part of the city without energy, but *no casualties* have been reported. According to unofficial sources, the *bomb* - allegedly detonated by *urban guerrilla commandos* - blew up a power tower in the northwestern part of <loc> San Salvador </loc> at <time> 0650 </time> (<time> 1250 </time> GMT).

Définition de trois sous-tâches:

- détection des entités
- détection des relations entre entités
- détection des événements

- **MUC** : extraire des entités nommées, soit des NP désignant, nommant un objet précis du monde.
- **ACE** : extraire des entités, nommées ou non. Il peut s'agir de NP, de GN, de pronoms. → détection des mentions d'une entité.

Recognition of entities, not just names. In the ACE entity detection and tracking (EDT) task, all mentions of an entity, whether a name, a description, or a pronoun, are to be found and collected into equivalence classes based on reference to the same entity. Therefore, practical co-reference resolution is fundamental. [DMP⁺ 04].

Typologie ACE

- 4 nouvelles catégories par rapport à MUC :
 - **Geo-political Entity** (gpe)
 - **Facility** (fac)
 - **Vehicle** (veh)
 - **Weapon** (wea)
- introduction d'une **hiérarchie** parmi les types et sous-types (pers = individus, groupes, indéfinis) ;
- **distinction** entre les expressions numériques (NUMEX) et les expressions temporelles (TIMEX).

N.B: il n'y a pas une mais des typologies ACE (bcp d'évolutions)

Typologie ACE

Types	Sous-types
PERS	individu, groupe, indéterminé
ORG	gouvernementales, commerciales, education, non gouvernementales, divertissement, media, religieuses, médical et sciences, sports,
GPE	continent, nation, état ou province, département ou région, villes, groupement de gpe, spécial, ainsi que des types comme pers, loc, org
LOC	adresses, frontières, objets astronomiques, plans d'eau, région géographique, région internationale, région autre
FAC	aéroports, usines, constructions, portion de construction
VEH	air, terre, eau, portions de véhicule, non spécifié
WEA	contondantes, explosives, coupantes, chimiques, biologiques, armes à feu, munitions, nucléaires, non spécifiés

Typologie ACE

Types	Exemples
FAC	L'aéroport Charles de Gaulle est grand.
GPE	Andorre se situe dans les montagnes.
LOC	M42 est une nébuleuse magnifique.
ORG	Le LDC est un laboratoire de recherche.
PER	Pierre roule sur la mousse avec la voiture.
VEH	les hélicoptères militaires ont ...; l' USS Alabama est un navire de ligne ...
WEA	des missiles sol-air ont été tirés...; le gaz sarin ...

NOMBREUSES AUTRES TYPOLOGIES S'INSPIRANT DE MUC ET ACE:

- **CoNLL** [TKSDM03]: inspiration MUC, ajout d'une catégorie MISC
- **HAREM** [SSCV06]: inspiration ACE, ajout de différentes catégories
 - **Idée** (*abstraccao*) : École (*escola*), Discipline (*disciplina*), Idéologie (*ideia*)
 - **Objet** (*obra*)
 - **Autre** (*variado*) : proche du *misc* de CoNLL
 - **Groupe**, appliqué à d'autres catégories : Titre (avec "groupe de titres" *grupocargo*), Personne (avec "groupe de personnes" *grupoind*) et Membre ("groupes de membres" *grupomembro*).
- **ESTER-2** [Est07]: encore plus de sous-types et traitement de l'imbrication

Typologie ESTER-2

Types	Sous-types
pers	pers.hum, pers.anim
fonc	fonc.pol fonc.mil fonc.admi fonc.rel fonc.ari
org	org.pol org.edu org.com org.non-profit org.div org.gsp
loc	loc.geo loc.admi loc.line loc.addr (+3) loc.fac
prod	prod.vehicule prod.award prod.art prod.doc
time	time.date (+ 2 abs et rel) time.hour (+ 2 abs et rel)
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

Imbrication d'entités (*nested entities*)

Au delà de la structuration en type et sous-types, il y a la **notion d'imbrication** :

- une entité peut en contenir une autre.
- *The <pers> president of <org> Ford </org> </pers>*

Structuration très utilisée dans des domaines de spécialité,
e.g. la typologie GENIA (domaine bio-médical) [KOTT03].

Vers une structuration fine des mentions

À la fin des années 2000, le programme Quaero définit une nouvelle typologie, utilisée dans la campagne ETAPE:

- inspiration ACE pour les catégories principales
- décomposition de la typologies (et de la tâche) en deux niveaux:
 1. caractérisation des types et sous-types (*type*)
 2. caractérisation des mots composants la mention (*composant*)

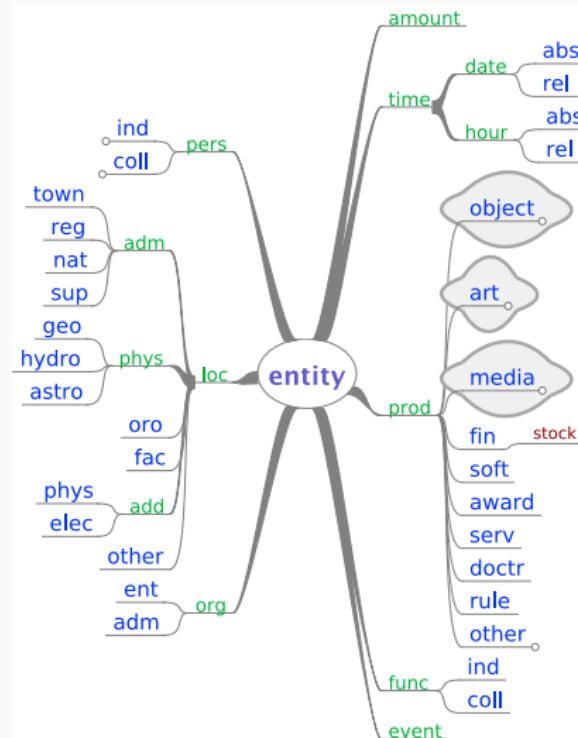
→ typologie hiérarchique et entités compositionnelles

Ref: [GRZ⁺11, RGZ11]

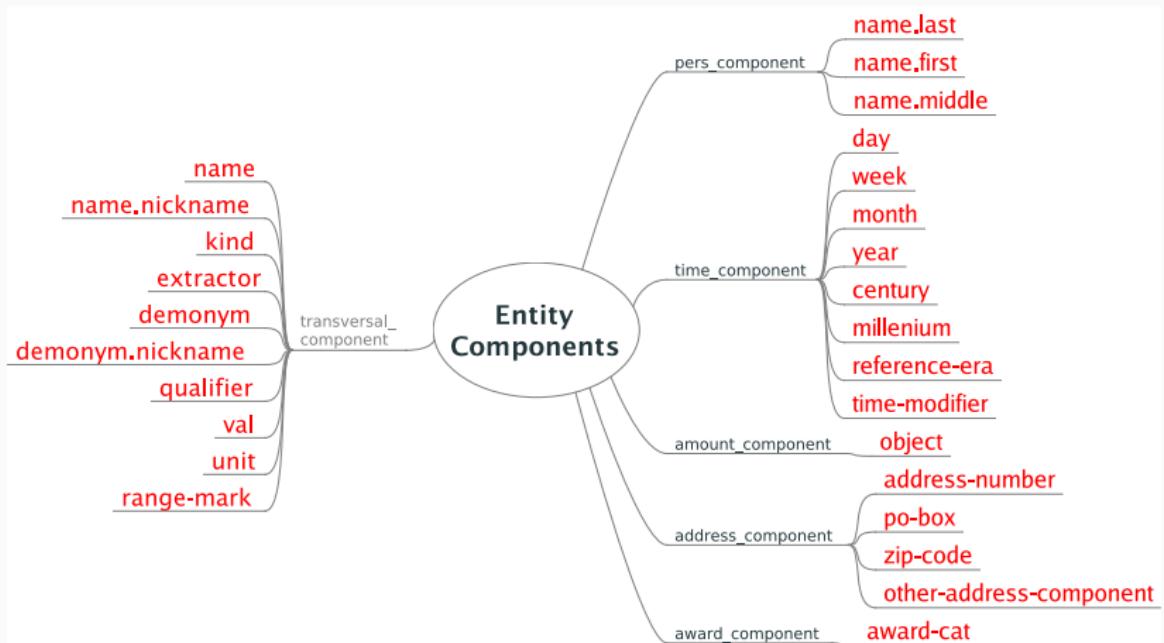
8 catégories principales

- **Person:** individual person, group of persons;
- **Location:** administrative location, physical location, facilities, oronyms, address;
- **Organization:** administration, service;
- **Time:** absolute and relative date, absolute and relative hour;
- **Amount;**
- **Product:** manufactured object, transportation route, financial products, doctrine, law, software, art, media, award;
- **Function:** individual function, collectivity of functions;

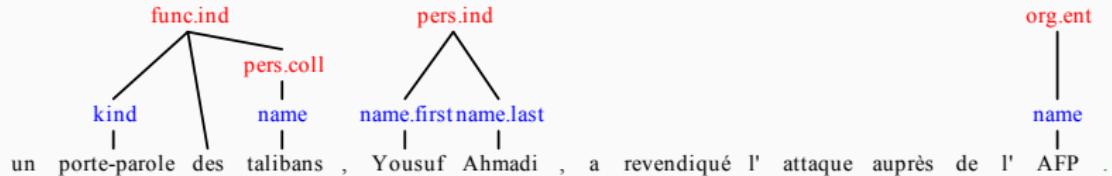
Typologie Quaero: sous-types



Typologie Quaero: composants d'entités



Quaero: composants d'entités



Les composants permettent :

- d'avoir, par compositionnalité, de nombreux types sans les multiplier
- d'aider au suivi et à la liaison, au moins intra-documents (l'usine Renault → l'usine)

Comparaison de typologies par l'exemple

MUC d'après le Bureau du recensement des LOC[Etats-Unis] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[2011] .

ACE d'après le ORG[Bureau du recensement des Etats-Unis] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[2011] .

ESTER d'après le ORG[Bureau du recensement des LOC[Etats-Unis]] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[2011] .

QUA d'après le ORG [name [Bureau du recensement] des LOC [name[Etats-Unis]]] , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE[year[2011]] .

Principales questions et points de divergences

1. **Gestion de la métonymie:** doit-on annoter le **sens en contexte** ou le **sens absolu** d'une entité?

e.g. *France* peut référer au pays, à l'organisation politique, à une équipe sportive.

2. **Gestion des entités imbriquées**

3. **Gestion des entités coordonnées**

Jacques et Bernadette Chirac: une ou deux entités?

Les différentes typologies apportent des réponses très variées.

Depuis 2009 : Text Analysis Conference Knowledge Base Population (TAC-KBP)

Pour une entité donnée, il importe de trouver de **nombreux attributs**.

E.g. pour une entité de type PERS:

- des **noms** : les autres noms que porte ou a porté cette personne (alias, faux noms, noms de scène, etc.) ;
- des **fonctions et activités** : ses emplois, ses occupations, etc. ;
- des **dates** (ou âge) : de naissance, de mort, des différents événements, son âge ;
- des **lieux** : les lieux en rapport avec des événements de sa vie comme la naissance, la mort bien entendu, mais aussi les différents emplois etc. ;
- des **personnes liées** : conjoint(e), enfants, autres membres de sa famille, etc. ;
- **autres informations** : les écoles et universités fréquentées, les pays visités, etc.

→ **un retour à la compréhension !**

Un retour à la compréhension

- La tâche reste très complexe malgré les progrès (importants)
 - en 2010 le meilleur système ne dépassait pas 0.30 de F-mesure, en 2014 le meilleur score était de 0.36 [SJ14]
- Les EN restent au cœur du processus :

*This year's slot filling evaluation represented an effort at continuity (...) It remains difficult to achieve F-measure higher than 30%. Reaching competitive performance on this task requires a fairly mature NLP system, such as **high-quality name tagging**, coreference resolution and syntactic analysis. [JGDG10]*

Pour une vue d'ensemble

<http://damien.nouvelets.net/resourcesen/typologies.html>

typologies: à retenir

- **Indispensable** à la tâche de NERC
- Fort héritage de **MUC** et **ACE**
- Grande **diversité** (plus de 20 typos inventoriées en 2016)...
- ...mais toujours **triade universelle** (person, organisation, lieu)
- Tendance à la **complexification** (imbrication, composants, knowledge base population)

Les typologies définissent le cadre d'action.

Elles sont indispensable à la création de *corpus*.

3. Resources

3.2 Corpus annotés

Corpus annoté

Un ensemble de documents textuels dont le texte est enrichi, lors d'une campagne d'annotation, par un marquage des entités nommées respectant une typologie donnée.

Typologies → Manuel d'annotation

- exemplification des catégories
- règles pour permettre à l'annotateur de faire des choix
- souvent, définition en parallèle de la typologie et de guide d'annotation

Campagne d'annotation

- à partir d'outils dédiés (BRATT, GLOZZ, WEBANNO)
- importance de la mesure de la qualité et de la cohérence des annotations
- publication du corpus avec des informations: sources, accord inter-annotateur, mesures utilisées, typologie et guide d'annotation.
- à faire avec soin: time and resource consuming !

Exemples de corpus français: ESTER 2

- campagne d'évaluation sur la transcription enrichie de la parole
- 6h de données orales transcrites manuellement et automatiquement
- 3 sous-corpus: apprentissage, dev, test
- ca. 90,000 entités
- format XML
- ELRAS0338

Exemples de corpus français: QUAERO

1. Corpus de parole transcrit

- ESTER 2 + ajouts d'autres documents
- utilisation de la typologie Quaero
- 2 sous-corpus: apprentissage & test
- ca 120,000 entités
- ELRA-S0349

2. Corpus de presse ancienne

- journaux du 19ème s. (sorties OCR)
- utilisation de la typologie Quaero
- 2 sous-corpus: apprentissage & test
- ca 150,000 entités
- ELRA-W0073

Exemples de corpus français: ETAPE

- campagne d'évaluation sur la transcription enrichie de la parole
- utilisation de la typologie Quaero
- 3 sous-corpus: apprentissage, dev, test
- ca. 30,000 entités
- format SGML

Vue d'ensemble des corpus existants

Recensement d'environ 160 corpus en 2016,
avec différent(e)s:

- langues (mais prédominance de l'anglais)
- domaines (mais prédominance du général)
- modalités (mais prédominance de l'écrit)
- typologies
- formats
- licenses
- méthodes de construction

Vue d'ensemble des corpus existants

<http://damien.nouvelets.net/resourcesen/corpora.html>

corpus: à retenir

- indispensable pour entraîner et évaluer
- lien étroit avec les typologies
- coûteux à élaborer
- dominance des langues d'Europe de l'Ouest et de la modalité écrite

De quoi a-t-on besoin pour traiter les entités nommées?

1. **Typologies**, pour définir d'un cadre sémantique
2. **Corpus annotés**, pour servir de référence (évaluation) et d'illustration
3. **Lexiques et bases de connaissances**, pour donner des informations sur les éléments à traiter (entraînement)

3. Resources

3.3 Lexiques et bases de connaissances

Lexiques et bases de connaissances

Objectif: fournir des informations relatives à des entités, en général ou dans des domaines de spécialité, sur lesquelles les systèmes automatiques peuvent s'appuyer afin de les reconnaître, les catégoriser et les désambiguïser.

2 types d'informations:

- **lexicales**, sur les unités composant les EN
- **encyclopédiques**, sur les référents des EN

Un élément central pour la reconnaissance et la classification des EN (mais évolution avec le deep learning).

Evolution importante de ce type de ressource depuis l'apparition de la tâche:

simple 'gazetteers' → encodage de plus en plus d'information

Encodent 2 types d'information:

- des **noms ou parties de noms d'entités** avec leurs types associés
→ directement utilisés pour reconnaître des unités équivalentes dans les textes
- des **mots amorces**, également avec leurs types associés. E.g. *Justin Trudeau*
→ des unités indiquant avec une forte probabilité la présence d'une entité d'un certain type. E.g. *Monsieur*

Constitution de bases lexicales

- forte **dépendance v-a-v du domaine** d'application
e.g. liste de mots amorces pour le domaine général vs. bio-médical
- défi 1: privilégier la **qualité** sur la quantité:
un petit nombre d'entrées suffit à reconnaître une majorité d'entités
- défi 2: se conformer à l'**évolution rapide** des entités nommés qui
sont une classe ouvert

ANNIE

- système d'extraction où il est possible de définir de lexiques
- partie intégrante de GATE (Sheffield University)
- lexiques utilisés par des règles ou un *lookup*
- lexiques propres à chaque système et chaque application
- dans le passé, difficile à partager

- base lexicale à large couverture pour l'anglais
- éloigné du monde de EN, peu de noms propres
- mais utile pour l'intégration de resources (e.g. Wikipedia + Wiktionary + WordNet)



- base lexicale multilingue pour EN
- représente des
 - entités ('pivots')
 - avec leurs formes de surface ('prolexèmes')
 - leurs types
 - et leurs relations (e.g. synonymie, méronymie)
 - v2.2: français (100k formes), anglais, polonais
- développée par l'Université François Rabelais de Tours

CNRTL Centre National de Ressources Textuelles et Lexicales

Ortolang Outils et Ressources pour un Traitement Optimisé de la LANGUE

cnrs atif

■ Accueil ■ Portail lexical ■ Corpus ■ Lexiques ■ Dictionnaires ■ Métalexicographie ■ Outils ■ Contact

■ Prolex

Le projet Prolex, piloté par le [Laboratoire d'informatique](#) (LI) de l'université François-Rabelais de Tours, a pour but de fournir, à la communauté du traitement automatique des langues (Tal), des connaissances sur les noms propres, qui constituent, à eux seuls, 10% des textes journalistiques. Ceci par la création d'une plate-forme technologique comprenant un dictionnaire électronique relationnel multilingue de noms propres (*Prolexbase*), des systèmes d'identification des noms propres et de leurs dérivés, des grammaires locales, etc.

La ressource Prolexbase est [un projet Tal](#) du LI, en collaboration avec :

- le laboratoire ligérien de linguistique ;
- l'université de Belgrade ;
- l'académie des sciences de Varsovie.

Ce projet a reçu le soutien :

- de l'action [Technolangue](#) du Ministère de l'Industrie (2003-2005) ;
- du programme d'action intégré Egide [Pavle-Savic](#) du Ministère des Affaires étrangères (2004-2005) ;
- du projet Feder Région Centre [Entités nommées et nommables](#) (2009-2010) ;
- du projet ERDF [Nekst](#) (2009-2014) ;
- du projet européen (CIP ICT-PSP) [Cesar](#) (2011-2013).

■ Prolexbase

La modélisation du domaine des noms propres définie dans le projet Prolex repose sur deux concepts centraux : le pivot et le prolexème. Le pivot ne représente pas le référent, mais un point de vue sur ce référent. Il possède dans chaque langue un concept spécifique, le prolexème, qui est une famille structurée de lexèmes. Autour d'eux, sont définis d'autres concepts et des relations (synonymie, méronymie, accessibilité, épynomie, etc.). Chaque pivot est en relation d'hyperonymie avec un type et une existence au sein de deux typologies.

Il n'est pas évident de définir la notion de nom propre. La plupart des définitions insistent sur le caractère unique de son référent et sur une sémantique et une syntaxe qui lui est propre. Nous avons choisi d'adopter le point de vue de (Jonasson, 1994) qui propose une définition plus large incluant ce qu'elle appelle les noms propres purs (nom de personne et nom de lieu) et les noms propres descriptifs qui résultent souvent de la composition d'un nom propre avec une expansion (Tour Eiffel, musée Rodin, etc.). Un nom propre descriptif peut être considéré comme une expression définie figée ou en cours de figement (Jardin des Plantes, Médecins sans frontières, etc.). Cette définition est assez proche de celle utilisée dans le domaine du Tal depuis la conférence MUC6.

Origine de la ressource LI (Université François-Rabelais de Tours)
Nature des données Lexique relationnel multilingue de noms propres
Soutiens institutionnels Action Technolangue du Ministère de l'Industrie
Programme d'action intégré Egide Pavle-Savic du Ministère des Affaires étrangères
Projet Feder Réseau Centre

<http://www.cnrtl.fr/lexiques/prolex/>

- toponymes et assimilés
- 7 millions d'entités et 10 millions d'entrées lexicales
- attributs: les coordonnées géospatiales, la population, le code postal, etc.
- attribution d'une URI à chaque entité
- 9 catégories principales (sous-divisées en 645 sous-catégories) :



9 catégories principales (sous-divisées en 645 sous-catégories) :

- entités administratives : pays, états, régions, etc.,
- hydronymes : fleuves, lacs, rivières, etc.,
- aires : parcs, réserves, champs, etc.,
- zones urbaines “peuplées” : villes, villages, etc.,
- routes : rues, autoroutes, etc.,
- bâtiments (en anglais, *spot*) : ponts, hôtels, musées, etc.,
- reliefs (en anglais, *hypsographic*) : montagnes, volcans, plages, etc.,
- entités sous-marines : bassins, lagunes, canaux, etc.,
- entités végétales : forêts, cultures, vignes, brousse, etc.

- un 'by-product' d'un système de veille médiatique:
7000 sources, 300k articles par jour, 70 langues, dont 21 avec traitement fin des entités nommées
- ca. 340,000 entités uniques (PERS et ORG)
- 1,7 million de variantes de noms (lexicalisations) dans 170 langues
- 32 millions relations cross-lingue, y compris entre différents jeux de caractères
- jusqu'à 400 variantes pour une entité



emm



10
Years
EMM
ESTD. 2008

Top Stories

UPDATED EVERY 10 MINUTES, 24 HOURS PER DAY.

Main Menu

- Top Stories
- 24 Hours Overview
- Events Detection
- Most Active Themes
- Help about EMM
- Overview
- Advanced Search
- Sources list
- Web Site Map

EU Focus

EU Policy Areas

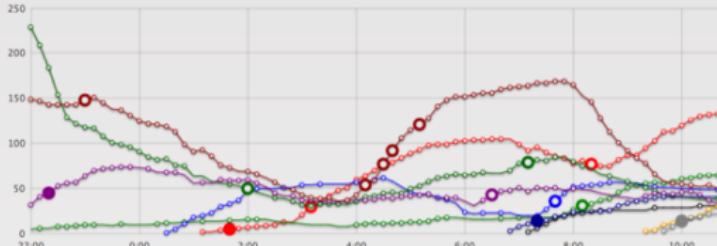
Themes

The World

Offices & Agencies

Current top 10 stories

Language: en Period: Jun 15, 2018 10:40 PM – Jun 16, 2018 10:40 AM



Multimillion-pound restoration hit by blaze at Mackintosh Building – fire chief Fire has caused "extensive" damage at Glasgow's famed Mackintosh Building...
enCA | Fire rages through historic Scottish school of art ↗
enCA Saturday, June 16, 2018 10:44:00 AM | Start : Jun 16, 2018 1:16:00 AM | Sources : 119 | Peak : 1 | Current rank : 1
Entities: Peter Capaldi[1]; Harry Potter[1]; James Bond[1]; Robbie Coltrane[1]; Nicola Sturgeon[1]; Paul Sweeney[1]; Simon Starling[1]; Martin Boyce[1]; Franz Ferdinand[1]; David Mundell[1]; Richard Wright[1]; Charles Rennie Mackintosh[2];
LONDON – Fire ripped through one of the world's top art schools, the Glasgow School of Art in Scotland, late on Friday. The historic building -- designed by Art Nouveau architect Charles Rennie Mackintosh -- was undergoing major restoration work following a blaze four years ago....
More articles...

Saudi-led forces seize airport in Yemen port city of Hodeida
expressindia Saturday, June 16, 2018 10:44:00 AM | Start : Jun 15, 2018 3:28:00 AM | Sources : 92 | Peak : 2 | Current rank : 2
Entities: Saudi Arabia[1]; Yemen[1]; Hodeida[1]; Port[1]; Express India[1];
Saudi-led forces seize airport in Yemen port city of Hodeida ↗
expressindia Saturday, June 16, 2018 10:21:00 AM CEST | Info [other]

Tools

Saturday, June 16, 2018
10:57:00 AM CEST

RSS | MAP

Facebook

subscribe | manage

info

Available on the App Store ANDROID APP ON Google play

Languages

Select top stories in other languages.

ar	bg	cs	da	de	el
en	es	et	fi	fr	hr
hu	it	lt	lv	mt	nl
pl	pt	ro	ru	sk	sl
sv	sw	tr	zh		

Show additional languages

Interface: en - English

Legend

Country Watch

The country most in the news at the moment.

<http://emm.newsbrief.eu>

Ehrmann, Rosset

Ecole thématique 'Big Data and Speech', Roscoff, Juillet 2018

96

Main Menu

[Top Stories](#)
[24 Hours Overview](#)
[Events Detection](#)
[Most Active Themes](#)
[Help about EMM](#)
[Overview](#)
[Advanced Search](#)
[Sources list](#)
[Web Site Map](#)

EU Focus

EU Policy Areas

Themes

The World

Offices & Agencies

Nicola Sturgeon

Last updated on 2018-02-21T08:07+0100.



ABOUT THIS IMAGE
LICENSING UNKNOWN
AUTHOR: THE SCOTTISH GOVERNMENT

Extracted quotes from

Nicola Sturgeon said : "not listened to, who is responsible and how are we going to ensure individuals are accountable?" [\[link\]](#)
thecourier Thursday, June 14, 2018 6:35:00 PM CEST

Nicola Sturgeon said : "Yesterday morning I was spending my time in two primary schools, as well as a secondary school and an early years centre. And I was talking to a range of primary school children including some five-year-olds. "I didn't meet any of them in tears, it didn't see any of them that looked crushed. What I saw were confident, bright enthusiastic young people - some of those were showing me computer coding and some were speaking Mandarin, that is how confident they were" [\[link\]](#)
bbc Thursday, June 14, 2018 4:32:00 PM CEST

Nicola Sturgeon said : "As local MSP for the Gorbals, I'm in close contact with NewGorbalsSHA who are on site." [\[link\]](#)
dailystar Thursday, June 14, 2018 1:53:00 PM CEST

Nicola Sturgeon said : "As local MSP for the Gorbals, I'm in close contact with NewGorbalsSHA who are on site." [\[link\]](#)
dailystar Thursday, June 14, 2018 12:21:00 PM CEST

Key Titles and Phrases (Last 30)

Names (Top 30)

KEY TITLES AND PHRASES	COUNT	LANG	LAST SEEN
minister	47.38%	EN	15/06/2018
leader	9.97%	EN	15/06/2018
première ministre écossaise	4.26%	FR	15/06/2018
ministre écossaise	3.87%	FR	15/06/2018
first minister of scotland	1.72%	EN	14/06/2018
minister of scotland	1.06%	EN	14/06/2018

Related entities (Top 30)

Associated entities (Top 30)

TYPE	ENTITY NAME	COUNT
EU	EU	7.23%
EU	Glasgow School	5.40%
EU	Charles Rennie Mackintosh	5.30%
EU	Ian Blackford	4.18%
EU	Theresa May	4.18%
EU	Paul Sweeney	3.97%

Articles published more than 12 hours ago

Tools

Saturday, June 16, 2018
10:58:00 AM CEST

Facebook

manage

Available on the [App Store](#) [Android APP ON Google play](#)

Languages

Select your languages

am	ar	az	be	bg	bs
ca	cs	da	de	el	en
eo	es	et	fa	fi	fr
ga	ha	he	hi	hr	hu
hy	id	is	it	ja	ka
km	ko	ku	ky	lb	lo
lt	lv	mik	ml	mt	nl
no	pap	pl	ps	pt	ro
ru	nw	si	sk	sl	sq
sr	sv	sw	ta	th	tr
uk	ur	vi	zh		
all					

Interface:
en - English

Legend

Explore Relations



Extracted quotes about

Adam Tomkins said (about Nicola Sturgeon) : "This is a remarkable report which exposes Nicola Sturgeon's secret Scotland. "People will see this report and

- publiée en format .txt en 2011 [add ref]
- RDF en 2016 [add ref]

Bio-médical: Meta-thesaurus UMLS

- *Unified Medical Language System*, développé par la *National Library of Medicine* [(MD09b)]
- unifie environ 200 terminologies, dont certaines sont multilingues
- e.g le thésaurus MeSH, dont la version française est développée par l'INSERM
- hiérarchies de concepts bio-médicaux, et relations entre ces derniers

The screenshot shows the U.S. National Library of Medicine (NLM) website. At the top, there's a blue header bar with the NIH logo, the text "U.S. National Library of Medicine", and a search bar. Below the header, a navigation bar includes links for "Databases", "Find, Read, Learn", "Explore NLM", "Research at NLM", and "NLM for You". On the right side of the header, there are links for "NLM Customer Support" and social media icons. The main content area has a yellow header "Unified Medical Language System® (UMLS®)". Below it, a breadcrumb trail shows "Home > Biomedical Research & Informatics > UMLS > Metathesaurus". The main text discusses the Metathesaurus as the largest component of the UMLS, organized by concept and linking names from over 200 vocabularies. It also mentions access via download, browser, or API, and provides a link to sign up on the UMLS Terminology Services (UTS) Web site. A "Release Information" section links to the UMLS Metathesaurus License Agreement and provides details about the release, including current and forthcoming releases. An "Education" section links to the UMLS Reference Manual. A sidebar on the left lists various UMLS components and their descriptions.

Unified Medical Language System® (UMLS®)

Home > Biomedical Research & Informatics > UMLS > Metathesaurus

Metathesaurus

The UMLS includes the Metathesaurus, the [Semantic Network](#), and the [SPECIALIST Lexicon and Lexical Tools](#). The Metathesaurus is the biggest component of the UMLS. It is a large biomedical thesaurus that is organized by concept, or meaning, and it links similar names for the same concept from nearly 200 different vocabularies. The Metathesaurus also identifies useful relationships between concepts and preserves the meanings, concept names, and relationships from each vocabulary.

You can access the Metathesaurus via [download](#), [browser](#), or [API](#).

Need an account? [Sign up](#) on the UMLS Terminology Services (UTS) Web site.

Release Information

See [Appendix 1](#) of the UMLS Metathesaurus License Agreement for a list of all vocabularies included in the latest version.

Follow these links for details about the release:

- [Current Release Documentation](#) – Release notes, bugs, license copy, metadata, and general statistics for the current Metathesaurus release.
- [Vocabulary Documentation](#) – Synopsis, contact information and statistics for each new and updated vocabulary in the current Metathesaurus release.
- [Forthcoming Releases](#) – Lists sources to be updated in future releases.
- [Release Documentation Archive](#) – Documentation dating from 2002AA.

Education

Read more about the Metathesaurus and its data files in [UMLS Reference Manual](#). Relevant chapters include:

- [Metathesaurus](#) – overview of the Metathesaurus and its organizational structure.
- [Rich Release Format \(RRF\) Data Files](#) – Descriptions of each file (MRCONSO.RRF, MRREL.RRF, etc.) including column names and samples records from each file
- [Original Release Format \(ORF\) Data Files](#) – Descriptions of each file (MRCON, MRREL, etc.) including column names and samples records from each file

[https://www.nlm.nih.gov/research/umls/knowledge_sources/
metathesaurus/](https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/)

Biologies

- bases lexicales sur les gènes et les protéines
- **UniProt**¹ (*Universal Protein Resource*) [C⁺¹⁰]
- avec des entrées validées manuellement:
SwissProt, 548 454 entrées
- et d'autres obtenues automatiquement:
TrEMBL, 47 452 313 entrées.

¹<http://www.uniprot.org/help/about>

Bilan sur les bases lexicales

- Ne sont pas toutes répertoriées ou publiées
- Au départ: décrire les réalisations linguistiques possibles d'entités
- Evolution: enrichissement avec d'autres informations, e.g. date de naissance d'une personne, population d'une ville, etc.
- 3 directions d'enrichissement:
 - couverture plus large
 - multilinguisme
 - information encyclopédique

→ structures de données plus complexes et volumineuses

Les bases de connaissances (survol rapide)

- Wikipedia (initiée en 2001)
 - utile pour extraire et intégrer des lexiques d'EN
 - constitution semi-automatique de corpus annotés
 - acquisition de relations entre entités
- DBpedia (équivalent RDF de Wikipedia)
- YAGO (Wikipedia, WordNet, plus infos spatiales et temporelles)
- BabelNet
- Wikidata
- OpenCyc (partie libre du Cyc), information de 'sens commun'

base lexicales et de connaissances: à retenir

- troisième pilier des ressources pour les EN
- information lexicales et sémantiques
- difficile à acquérir, représenter, stocker jusqu'au milieu 2000
- aujourd'hui: explosion d'information, principalement pour le domaine général



(20-30min break?)

SESSION 2

4. Reconnaissance et classification

Traitement des EN - rappel des tâches

1. **reconnaissance**: détecter, repérer des entités nommées dans les flux textuels (on pose les frontières dans le texte)
2. **classification**: catégoriser les éléments reconnus selon des catégories sémantiques pré-définies (on affecte un type)
3. **désambiguisation/liaison**: lier les mentions d'entités à une référence unique (on lie à une référence)
4. **extraction de relation**: découvrir des relations entre entités (*father-of, born-in, alma mater*)

Objectifs

Construire des systèmes logiciels qui effectuent ces tâches de manière automatique.

Exigences:

- **qualité**: ne pas faire trop d'erreurs
- **exhaustivité**: ne pas manquer trop d'entités
- **robustesse**: ne pas échouer face à des cas non canoniques

En pratique:

- difficile de répondre à ces 3 exigences simultanément
- recherche du **meilleur compromis** en fonction des ressources et de l'application.

Matériaux de base

texte, représenté comme une **structure linéaire**, i.e. une séquence de mots pouvant être découpée en **segments**

Objectif

indices portés par les segments \iff présence d'une entité nommée
 \iff d'une certaine catégorie
 \iff référant à une certaine entité

attention: pas de correspondance systématique entre un ensemble de propriétés et une classe d'entités

Vue d'ensemble des méthodes

- méthodes symboliques à base d'automates
- méthodes numériques (heuristiques ou pondérations)
- méthodes statistiques, *i.e.* apprentissage automatique, supervisé, semi-supervisé ou non supervisé

4. Reconnaissance et classification

4.1 Indices

Représentation du texte

La **représentation des textes** comme séquences de mots donne 2 niveaux de granularité:

- les **caractères**, qui forment un mot
- les **mots**, qui composent une séquence (un texte)

Les **indices** peuvent être caractérisés au niveau:

- des caractères: **indices morphologiques**
- des mots eux-mêmes: **indices lexicaux**
- de la séquence de mots: **indices contextuels**

Indices morphologiques

C'est à dire:

- les caractères qui constituent les mots
- l'existence de différentes classes de caractères
- l'existence de régularités dans l'agencement des caractères, dont certaines sont utiles pour les EN

La majuscule

- très utilisée dans les jeux de caractères occidentaux pour annoncer un nom propre
- très facile à tester pour détecter des EN

MAIS

- majuscules également présentes en début de phrase, dans les sigles
- utilisée pour les noms communs en allemand
- n'aide pas à la classification
- la notion de casse est minoritaire en orient (n'existe pas en chinois, hindi, arabe, etc.)

Régularités socio-culturelles

- le suffixe *-ville* ou le préfixe *Saint-* en français
- suffixes réguliers pour les noms de personnes
 - en russe, le suffixe *-vitch*
 - en suédois, le suffixe *-sson*
 - en islandais, *-dóttir*
 - en Afrique du nord, les préfixes *Ben-* ou *Aït-*
 - en japonais, le suffixe *-san*
- mots issus de conventions ou de normes :
Inc. en anglais, *S.A.* en français, *GmbH* en allemand pour les ORG

Présence de nombres

- les dates, montants ou mesures, contiennent typiquement des nombres (écrits en chiffres ou lettres)
- chiffres plus ou moins soudés (*10 000, quatre-vingt-treize, cent dix-huit*) et/ou avec des caractères spécifiques (*10,38, 24/03*).
- mélange de caractères numériques et alphabétiques (*100km, 10h30, etc.*).
- sigles et d'abréviations (*A380, ISO-9000, Canon EOS 70D*).

- **En général:** attribution d'*étiquettes* aux mots pour distinguer e.g. les noms propres, chiffres, adverbes, déterminants, noms communs, etc.
- **Appliquée aux EN:**
utilisation de motifs morphologiques et des motifs morphologiques résumés:
i.e. quelles combinaisons de classes de caractères (alphabétiques, ponctuations, chiffres) sont utilisées pour former quels mots
→ mécanismes de reconnaissance à l'aide d'*expressions régulières*

Bilan indices morphologiques

Détection d'entités caractérisées par des régularités, pour un certain nombre de langues.

Mais ils restent insuffisants:

- ne couvrent que les formes très normées d'EN
- permettent de *déetecter*, mais pas de *catégoriser*

Principe: confronter les textes à des listes d'entités de ou de composants d'entités.

- mécanisme très précis si entrées lexicales contrôlées
- lexiques souvent organisés selon les types d'EN ou le degré d'ambiguïté
 - e.g. *Hollande vs. Obama*
- attention à trouver le bon compromis entre quantité et efficacité

En pratique: les algorithmes retournent la liste des segments correspondant aux occurrences des entrées des lexiques.

Attention: mécanisme est ambigu, plusieurs entrées peuvent être retournées.

Aujourd'hui, François Hollande a rencontré Obama à Washington.

- la Personne *François Hollande*
- le Pays *Hollande*
- la Personne *Obama*
- la Personne ou le Lieu *Washington*.

Défi: les EN sont une classe **ouverte**, impossible d'être exhaustif

- de nouveaux noms propres se créent continuellement
[McD96, Fri02]
- certaines parties de descriptions définies sont substituables

→ tenir à jour un lexique d'EN avec toutes leurs formes exactes et leurs variations est une tâche coûteuse et complexe.

Souvent, **prise en compte conjointe** d'indices morphologiques et lexicaux:

- **Personnes** : le premier mot est un prénom, le second un nom propre
- **Dates** : le premier et le dernier mot sont composés de chiffres, le second mot fait partie de la liste des noms de mois (*5 juillet 2012*)
- **Lieux** : contient *sur* ou *en* suivi d'un nom de cours d'eau (*Montlouis sur Loire*)
- etc.

L'examen des mots qui composent les entités ne dit pas tout.
Les indices morpho et lexicaux peuvent être absents ou ambigus.

→ besoin d'indices complémentaires, à proximité:

- **contexte local:** mots qui précèdent ou suivent l'entité.
- **contexte global:** phrase, phrases proches, paragraphe, document.

Importance des indices contextuels

1. *Il a vu Hollande à la télévision.*
2. *Son voyage en Hollande s'est bien passé.*
3. *Il a acheté une Renault Clio.*
4. *La muse Clio chante le passé des hommes et des cités.*
5. *Je me documente sur Washington pour mon travail.*

La graphie de l'entité étant identique, seul un appel au contexte permet de classifier.

Facile et intuitif pour l'humain, plus compliqué pour une machine.

- traitement appliqué non aux mots, mais à des pans de textes
→ coût computationnel plus élevé
- souvent besoin de s'appuyer sur des analyses préliminaires:
syntaxique, coréférences, thématique du document
- plus économique: sélectionner, *a priori* ou *a posteriori* les indices contextuels les plus discriminants et leurs combinaisons
- analyses en contexte importante lorsque la typologie des EN est fine

indices: à retenir

- de nature morphologique, lexicale ou contextuelle
- possibilité d'indices composites, par conjonction ou disjonction
- ce sont les 'ingrédients' des systèmes automatiques de traitements d'EN

4. Reconnaissance et classification

4.2 Approches symboliques

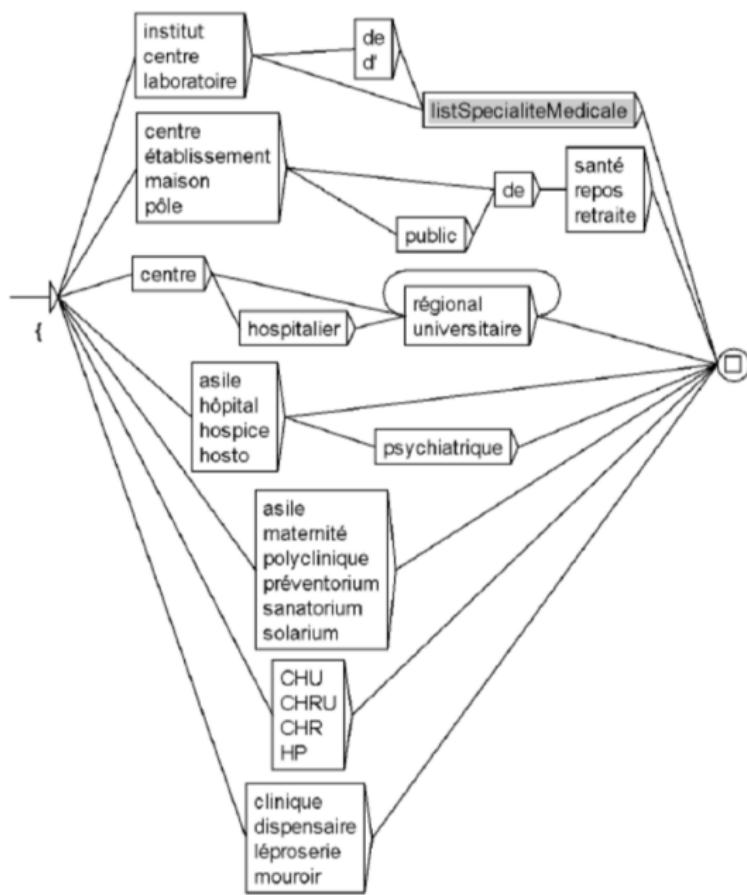
Techniques à base d'automates

- **Objectif:** insertion de balises dans les textes indiquant où se trouvent les ENs
- **Principe:** conception de *règles* formant un *grammaire locale*
- **Réalisation:** utilisation de transducteurs
chaque règle est associée à un diagramme d'états où les nœuds représentent les états de l'automate et les arêtes les transitions
- De nombreuses **boîtes à outils**:
 - GATE
 - LingPipe
 - NooJ
 - OpenNLP
 - OpenCalais
 - Unitex

- des éléments sont indiqués au sein des nœuds
- les noeuds sont agencés de manière à reconnaître des expressions linguistiques
- les transitions sont réalisées par présence d'indices (morphologiques, lexicaux, internes ou externes)
- plusieurs transitions sont réalisées par juxtaposition de nœuds
- l'automate ne reconnaît une expression linguistique que s'il existe un chemin depuis le nœud initial (à gauche) jusqu'au nœud final (à droite).

Objectif : contraindre correctement l'automate, afin qu'il reconnaisse toutes les expressions linguistiques souhaitées, et aucune autre.

Automates



Possibilité d'avoir des prétraitements:
segmentation en mots, en phrases, étiquetage morphosyntaxique.

→ indices supplémentaires fort utiles,
mais qui impactent les performances si bruités.

Basculement vers les approches statistiques

Au début des années 2000, grâce à la mise à disposition de jeux de données volumineux.

Mais les approches symboliques sont toujours présentes:

- combinées avec des méthodes statistiques
- prédominent pour les langues ou les typologies sans corpus de données suffisants
- gardent l'avantage pour le contrôle et de l'ingénierie: plus compréhensibles, modulables, possibilités de réglages fins.
- majoritaires dans le milieu industriel.

4. Reconnaissance et classification

4.3 Modèles guidés par les données et apprentissage

Le paradigme de l'apprentissage automatique

Objectif: déterminer les paramètres d'un modèle à partir de données,
d'où le terme *apprentissage*

Ces paramètres et ce modèle sont ensuite utilisés pour prendre les décisions les plus probables (ou vraisemblables) sur de nouvelles données à traiter.

Il s'agit, simultanément, de spécifier le modèle et de généraliser les données.

Le paradigme de l'apprentissage automatique

A partir des années 1960, émergent les modèles connexionnistes (perceptron, réseaux de neurones), qui mettent en relation des propriétés sur les objets modélisés.

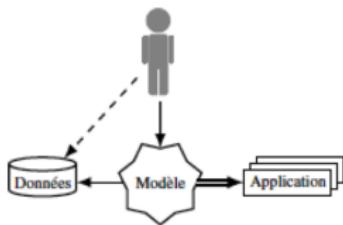
Puis les modèles markoviens (modèles de Markov à états cachés), qui simulent des processus stochastiques.

→ remise en cause du principe déterministe des automates et la manière dont sont élaborés les systèmes.

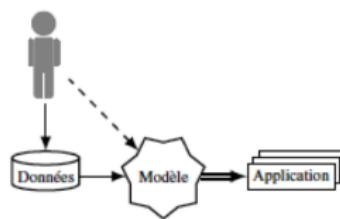
Le paradigme de l'apprentissage automatique

Systèmes symboliques: le concepteur du système interagit majoritairement avec le modèle (l'automate), et n'utilise les données que pour visualiser ou d'évaluer.

Systèmes guidés par les données: le concepteur agit sur les données, la structure du modèle est prédéfinie et rigide et les paramètres ajustés automatiquement à partir des données.



Système symbolique



Système guidé par les données

Le rôle des données

- format
- quantité
- qualité

→ modèles plus ou moins précis, couvrants et robustes

Formats tabulaires. Après une segmentation en mots, une étiquette est associée à chaque mot:

- catégorie d'entité nommée (PERS, ORG, etc.)
- une catégorie spéciale, Ø (comme *Outside*)

Problème: entités polylexicales vs entités nommées contigües de même type. e.g. *Paris Brest*

Solution: catégories éclatées

- Format **BIO** ($2N + 1$ classes différentes):
 - mot qui commence une nouvelle entité (PERS-B pour *Begin*)
 - mot qui prolonge une entité polylexicale (PERS-I pour *Inside*)
- Format **BILOU** ($4N + 1$):
distingue aussi les entités composées d'un seul mot ou les derniers mots des entités: *Begin*, *Inside*, *Last*, *Outside*, *Unique*
→ de meilleures performances si données annotées suffisantes

Exemple

En 2008 Hollande prend un vol Rio de Janeiro Los angeles

Balises	En <DATE> 2008 </DATE> <PERS> Hollande </PERS> prend un vol <LOC> Rio de Janeiro </LOC> <LOC> Los angeles </LOC>
BIO	En 2008/DATE-B Hollande/PERS-B prend/0 un/0 vol/0 Rio/LOC-B de/LOC-I Janeiro/LOC-I Los/LOC-B angeles/LOC-I
BILOU	En 2008/DATE-U Hollande/PERS-U prend/0 un/0 vol/0 Rio/LOC-B de/LOC-I Janeiro/LOC-L Los/LOC-B angeles/LOC-L

Déterminer la classe d'un mot à partir de la classe qui lui est majoritairement associée dans le corpus d'apprentissage.

Formulation à l'aide de probabilités:

- fréquence du mot $F(m)$
- fréquence d'une étiquette $F(e)$
- fréquence de la présence jointe du mot et de l'étiquette $F(m, e)$

La formule de Bayes et l'estimation statistique permettent de calculer la probabilité d'une étiquette sachant le mot :

$$P(E_i = e|M_i = m) = \frac{P(M_i = m, E_i = e)}{P(M_i = m)} = \frac{F(e, m)}{F(m)}$$

Probabilité d'une étiquettes pour un mot donnée =
ratio entre la fréquence dans le corpus annoté du mot avec une étiquette
et la fréquence dans ce même corpus du mot (quelque soit l'étiquette)

Modèles par classes majoritaires

Pour une séquence de mots et d'étiquettes (hypothèse d'indépendance entre les mots):

$$P(E_1, E_2 \dots E_n | M_1, M_2 \dots M_n) = \prod_{i=1}^n P(E_i | M_i) = \prod_{i=1}^n \frac{F(e, m)}{F(m)}$$

Puis: sélectionner la suite d'étiquettes qui, en fonction des mots de l'énoncé, maximise cette probabilité.

Complexité restreinte ici: choix de l'étiquette la plus probable pour chaque mot.

Modèles par classes majoritaires

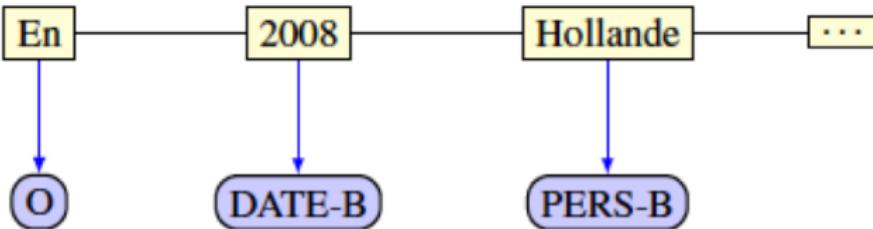


Figure 4.3. Modèle par classes majoritaires

(l'orientation des flèches indique quelles dépendances sont prises en compte par le modèle)

Modèles à décisions contextuelles (HMM)

Objectif: tenir compte de la vraisemblance d'**étiquettes contiguës**

François Hollande

- *Hollande*: Lieu ou Personne ?
- *François*, annoté comme Personne, peut conditionner l'annotation du mot *Hollande*

Modèles à décisions contextuelles (HMM)

Option: modèles génératifs comme les modèles de Markov à états cachés.

Calcul des probabilités inversé : déterminer, pour une suite d'étiquettes, la probabilité qu'elle génère un texte donné.

$$P(M_1, M_2 \dots M_n | E_1, E_2 \dots E_n) = \prod_{i=1}^n P(M_i | E_i) * P(E_i | E_{i-1})$$

Soit le produit des probabilités de génération $P(M_i | E_i)$ et de transition $P(E_i | E_{i-1})$.

Modèles à décisions contextuelles (HMM)

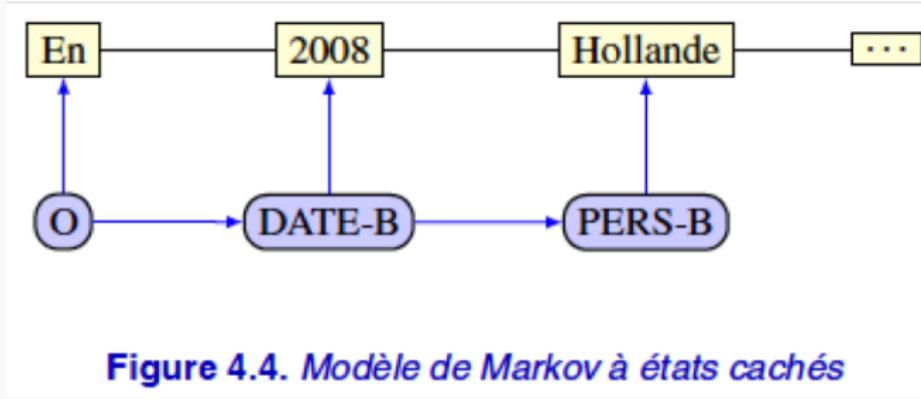
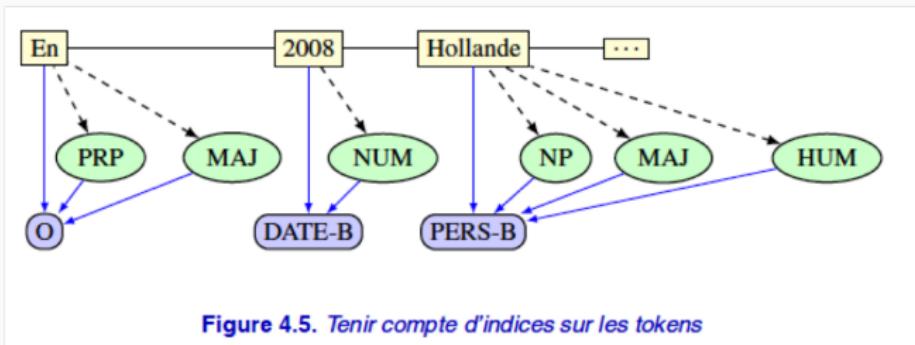


Figure 4.4. Modèle de Markov à états cachés

Décisions non indépendantes : la solution la plus vraisemblable est choisie en fonction des étiquettes préalablement choisies.

Modèles utilisant des indices multiples (softmax, MaxEnt)

Objectif: considérer plus d'indices que les mots, i.e. prendre en compte la morphologie, les indices lexicaux, le contexte, etc.



Modèles utilisant des indices multiples (softmax, MaxEnt)

Utilisation d'une fonction dédiée G , qui utilise des statistiques sur les indices $F_{i1} \dots F_{ik}$ dans un calcul, ainsi qu'à une normalisation parmi les T types d'entités nommées possibles.

$$P(E_i = e | M_i = m, F_{i1} = f_1 \dots F_{ik} = f_k) = \frac{G(e, m, f_1 \dots f_k)}{\sum_{t \in T} G(t, m, f_1 \dots f_k)}$$

Il existe différentes manières de définir la fonction G .

Champs markoviens conditionnels (CRF)

Les CRF (*Conditional Random Fields* ou champs markoviens conditionnels) combinent les deux aspects précédents :

- tenir compte du contexte pour prendre des décisions
(une décision sur un mot influence la décision pour le mot suivant)
- tenir compte de multiples indices
(analyses en prétraitements)

Modèle qui obtient de très bonnes performances pour la reconnaissance d'EN.

Champs markoviens conditionnels (CRF)

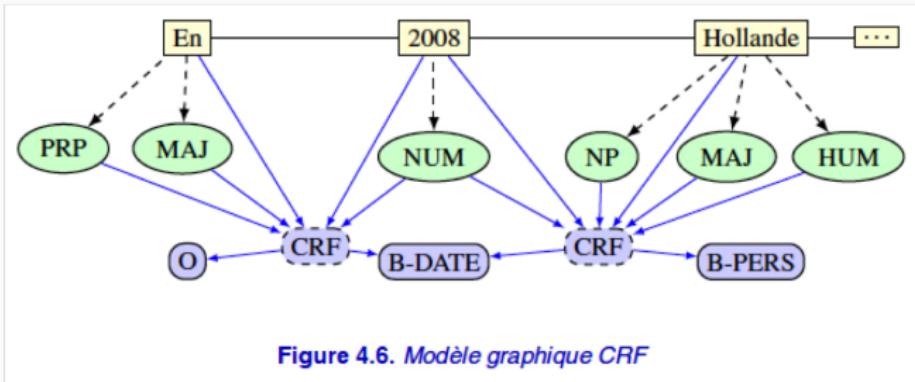


Figure 4.6. Modèle graphique CRF

$$G(e, m, f_1 \dots f_k) = \exp \left(\sum_{p=1}^k \alpha_{ep} * f_p \right)$$

Réseaux de neurones profonds

- Illinois NE Tagger

http:

//cogcomp.cs.illinois.edu/page/software/_view/NETagger

- Stanford NE tagger

<http://nlp.stanford.edu/software/CRF-NER.shtml>

- ...

Reconnaissance et classification d'en: à retenir

- possibilité d'utiliser de nombreux indices
- via des méthodes diverses qui peuvent être combinées.
- si plus d'indices, alors complexité grandissante et besoin de plus de données annotées
- importance de la sélection d'indices



(10min)

5. Liaison

- nous savons reconnaître et catégoriser des segments textuels:
des *mentions* d'entités qui font référence à un objet du monde.
 - ce qu'il reste à faire: établir le lien entre les mentions et les objets auxquels elles réfèrent
- objectif: désambiguïsation, résolution, liaison

Des mentions aux référents

- **Catégoriser n'est pas désambiguiser:**

G. Bush et F. Mitterrand sont des PERSON

Mais lequel des 2 réfère au *43ème président des États-Unis*?

- **Le problème des homonymes:**

F. Mitterrand est une PERSON

Mais *François Mitterrand* ou *Frédéric Mitterrand* ?

Bush est une PERSON

Mais *G. W. Bush* ou *G. Bush* ?

- **Le problèmes des variantes:**

Jean-Claudem Junckerem, Juncker, Jean-Cluade Juncker et le président de la Commission Européenne réfèrent-elles à la même entité?

- **Résolution de co-référence:**
au sein d'un même document, identifier que *Frédéric Mitterrand*, *Mitterrand*, *FM* ont le même référent (quel qu'il soit)
- **Clustering de mentions:**
pour une collection de documents, identifier que *Frédéric Mitterrand*, *Mitterrand*, *FM* ont le même référent (avec ou sans référentiel)
- **Liaison d'entités:**
étant donnés des documents, identifier les mentions d'entités et lier chacune d'elles à un référent d'une base de connaissances

- forte utilisation de Wikipédia ('*wikification*') et/ou DBpedia
- lorsque le référent d'une mention est absent de la base → NIL
Non trivial: possibilités de mentions dont le référent est absent de la base, mais dont un homonyme y est présent.
→ vers la population de bases de connaissances (cf. TAC-KBP)

Formalisation

Etant donnés:

- l'ensemble des mentions dans des textes
- l'ensemble des entités référencées dans une base

la liaison est une application depuis le domaine des mentions vers le domaine des référents.

- ni injective: plusieurs mentions peuvent être associés à un référent
- ni surjective: tous les référents n'ont pas à être liés

La liaison peut s'appuyer sur une *reconnaissance* des EN : restreindre les référents potentiels selon le type facilite les choses (e.g. *Washington* pour le gouvernement des États-Unis).

Formalisation

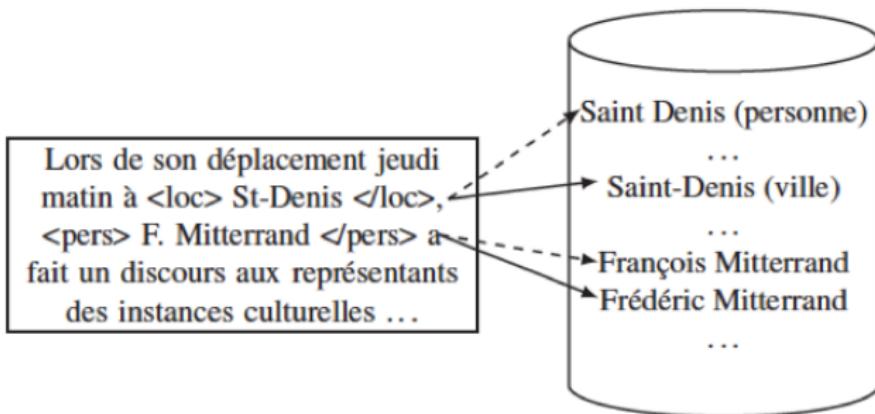


Figure 5.1. Mentions du texte et références de la base de connaissances

Etapes du processus de liason

1. Recherche des mentions d'EN dans les textes
2. Selection de candidats dans la base
3. Liaison

Recherche des mentions

1. utilisation de systèmes de reconnaissances d'EN (cf. section précédente) exposés
2. large collection de mentions par simples heuristiques ou par lookup des noms de la base

N.B.: une collection 'grossière' de mentions n'est pas préjudiciable: l'étape suivante fait un filtre.

Selection de candidats

Confronter chaque mention à tous les référents est coûteux et peu efficace.

→ on sélectionne, pour chaque mention, les référents *candidats* à la liaison.

Approche:

- prise en compte des mots des mentions du côté du texte (ou les formes de surface)
- prise en compte des variantes (ou *dénominations, lexicalisations*) présentes dans la base pour les référents.

Récupération de dénominations dans la base

Pour Wikipedia:

Élément générique	Donnée Wikipédia
Nom du référent	Titre de l'article
Autres appellations indiquées	Texte en gras du premier paragraphe
Synonymes	Pages de redirection
Textes des pointeurs	Liens internes

Table 1: Extraction de variantes des référents dans une ressource

Selection de candidats

Donc: Pour une mention donnée, récupération des référents qui ont une variante correspondant aux mots de la mention.

Exemple: la mention *F. Mitterrand* aura comme candidats toutes les personnes dont le prénom commence par un *F* et dont le nom de famille est *Mitterrand*

Selection de candidats

Traitements de surface pour ne pas être trop contraignant:

- insensibilité à la casse (majuscules / minuscules) ;
- variantes liées aux jeux de caractères (diacritiques, liaisons, ponctuations, etc.) ;
- suppression des mots-outils ;
- suppression d'éléments entre parenthèses ;
- génération automatique d'acronymes à partir des formes de surface ;
- etc.

Enjeu: multiplier les variantes linguistiques pour chaque référent afin de ne pas manquer des mentions.

Remarque

Les 2 étapes de collection de mentions et de sélection de candidats peuvent être remplacées par une reconnaissance d'EN performante privilégiant le rappel sur la précision.

Etape la plus importante.

Objectif: associer à chaque mention le candidat le plus vraisemblable (ou NIL)

Méthode: prises en compte d'**indices** du côté de la mention et du côté du référent, et **calcul** d'une distance.

- Coté mentions:
 - mots de la mention
 - contexte immédiat de la mention
 - texte du document
- Coté référents:
 - éléments textuels de sa description (titre, synonymes, résumé, article)
 - autres propriétés attribuées au référent (infobox)
 - entités et concepts associées (internes et externes)

Beaucoup d'informations disponibles, il faut choisir les plus pertinentes.

- **Indices textuels**

→ possibilité de calculer une distance cosinus

Dans quels cas cela marche-t-il bien?:

- *Washington* LIEU vs. PERSONNE
- *Karl Marx versus Thierry Marx*
- *George H. W. Bush versus George W. Bush*

- **Indices structurels:**

→ sélection du référent le plus populaire selon un critère.

- pour tout référent, le nombre de liens pointant vers sa page
- pour une ville, son nombre d'habitants

Attention: non prise en compte des mentions: résultats toujours identiques.

Performances

- **Premiers travaux exploratoires**

briques élémentaires, définition des différentes sous-tâches [BP06, Cuc07]

- **TAC 2009**

- 82.2% d'accuracy (meilleur système basé sur DBpedia)
 - dont 76.5% sur les entités à lier et 86.4% pour les entités

- **TAC 2011**

- 85% et 90% selon les données considérées
 - Performances des humains: autour de 90%

→ Bonnes performances sur des textes génériques en anglais.
Références:[MD09a, JGDG10, VBKR⁺09, DMR⁺10, RMD13, JGDG10].

liaison: à retenir

- établissement de liens entre les *textes* et des *bases de connaissances*
- essor de la tâches à partir de 2007
- relatives bonnes performances pour l'anglais
- Wikipédia est la principale, voire unique, base de connaissances

Possibilité de s'appuyer sur les entités reconnues et liées pour mieux comprendre les textes

→ bénéficiaires: reconnaissance de la parole, la traduction automatique, le résumé automatique, recherche d'information

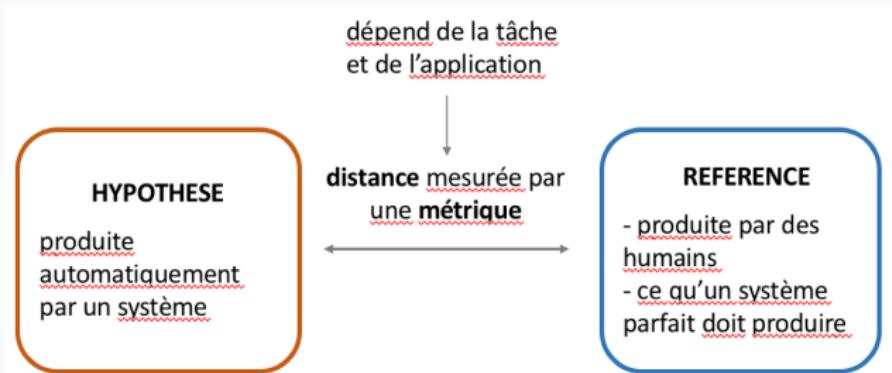
6. Evaluation

6. Evaluation

6.1 Introduction

- Première formalisation de la procédure d'évaluation: MUC3 [Sun91]
- Motivation: avoir des éléments de comparaison stables et effectifs entre hypothèses et références

Protocole d'évaluation



Objectif: mesurer à quel point le système trouve les “bonnes réponses”

Quelle “bonnes réponse”?

- traduction ou le résumé automatique : bonne réponses multiples
- EN: on peut supposer une seule et unique “bonne réponse”

- **Transparence:** “règles du jeu” connues par tous
- **Coût:** réduit par rapport à une évaluation manuelle pour chaque hypothèse des systèmes ;
- **Reproductibilité:** réutilisation au delà des campagnes permettant une comparaison des résultats dans la production scientifique

Ce qu'il faut pour évaluer

1. Une **métrique** mesurant la distance entre une référence et une hypothèse ;
2. Un **algorithme d'alignement** de la référence et de l'hypothèse.

6. Evaluation

6.2 Les mesures classiques

Précision

Ratio entre le nombre de **réponses correctes** et toutes les **réponses données** par un système

$$P = \frac{C}{C + S + I} \quad (1)$$

- C : nombre d'objets **corrects** dans l'hypothèse;
- I : nombre d'**insertions** par le système ;
- S : nombre de **substitutions** par le système (entités mal typées).
- soit $C + S + I$: nombre total d'objets dans l'hypothèse.

Rappel

Ratio entre le nombre de **réponses correctes** et le nombre des **réponses attendues** (i.e. présentes dans la référence)

$$R = \frac{C}{C + S + D} \tag{2}$$

- D : nombre total d'**omission** (*deletions*) opérées par le système (entités non détectées) ;
- $C + S + D$: nombre total d'objets dans la référence.

Exemple 1

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

- Precision = $\frac{2}{3} = 0,67$
- Rappel = $\frac{2}{2} = 1$

→ ici HYP1 produit du **bruit**

Exemple 2

REF: <pers> Bertrand Delanoë </pers> a été élu maire de
<loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> a été élu maire de
Paris

- Precision = 1
- Rappel = $\frac{1}{2} = 0.5$

→ HYP2 produit du **silence**

- La précision tient compte des **insertions** et **substitutions**
- Le rappel tient compte des **omissions**

Comment combiner les 2 en une seule mesure?

F-mesure, définie comme la **moyenne harmonique entre Précision et Rappel**:

$$F = (1 + \beta^2) \times \frac{P \times R}{\beta^2 P + R} \quad (3)$$

Où β est un **poids** permettant d'ajuster l'importance de P ou R (si 1, égale importance).

Exemples

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> a été élu maire de Paris

$$F(HYP1) = (1 + 1^2) \times \frac{0,67 \times 1}{1^2 \times 0,67 + 1} = 0,80 \quad (4)$$

$$F(HYP2) = (1 + 1^2) \times \frac{1 \times 0,5}{1^2 \times 1 + 0,5} = 0,67 \quad (5)$$

Inconvénients des mesures classiques

- Fusionner P et R minimise le poids des erreurs d'insertion et d'omission par rapport aux erreurs de substitution, quel que soit β [MKS99]
- Avec les typologies fines et complexes, besoin d'une métrique différenciant les erreurs.

Différents types d'erreur

REF: the <pers.ind> president of Ford </pers.ind>

HYP1 : the <pers.ind> president </pers.ind> of Ford
→ erreur de frontière

HYP2 : the <pers.coll> president of Ford </pers.coll>
→ erreur de sous-type

HYP3 : the <pers.coll> president </pers.coll> of Ford
→ erreur de sous-type et de frontière.

6. Evaluation

6.3 Les mesures basées sur le décompte d'erreurs

ERR, Error Per Response

- définie lors de MUC5 [CS93]
- inspirée du taux d'erreurs mots (WER pour *Word Error Rate*) en RAP [Pal85]
- mesure des erreurs: plus le taux est bas, mieux c'est.

$$ERR = \frac{S + D + I}{C + S + D + I} \quad (6)$$

ERR: exemples

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP1: <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> a été élu maire de Paris

$$ERR(HYP1) = \frac{0 + 0 + 1}{2 + 0 + 0 + 1} = \frac{1}{3}$$

$$ERR(HYP2) = \frac{0 + 1 + 0}{1 + 0 + 1 + 0} = \frac{1}{3}$$

Le poids des insertions est moins important que celui des substitutions et des omissions[MKSW99].

Une augmentation de I provoque une augmentation de ERR moins importante qu'une augmentation de $S + D$.

$$ERR = \frac{S + D + I}{N + I} \quad (7)$$

Avec N = nombre d'entités dans la référence.

ERR: problème

Pour $N = 100$, $S + D = 10$, $I = 10$, on a:

$$ERR = \frac{10 + 10}{100 + 10} = \frac{20}{110}$$

Si on augmente $S + D$ de 10:

$$ERR = \frac{20 + 10}{100 + 10} = \frac{30}{110} = 0,27$$

Si on augmente I de 10:

$$ERR = \frac{10 + 20}{100 + 20} = \frac{30}{120} = 0,25$$

De plus, avoir I dans le dénominateur rend les résultats non comparables.

SER: Slor Error Rate

- proposée par [MKSW99]
- identique au WER utilisé en RAP
- utilisée lors de ACE, ESTER-2, QUAERO et ETAPE
- suppression du nombre d'insertion (I) du dénominateur:

$$SER = \frac{S + D + I}{C + D + S} = \frac{S + D + I}{R} \quad (8)$$

où R = nombre total d'entités de la référence.

Possibilité d'affiner l'importance relative des erreurs:

$$SER = \frac{\alpha_1 S_t + \alpha_2 S_f + \beta D + \gamma I}{R} \quad (9)$$

- S_t et S_f : nombre total de substitution de type et de frontières ;
- D et I: nombre total d'omissions et insertions ;
- α_1 α_2 β et γ : poids affectées à chaque catégorie d'erreur.

SER: inadapté aux imbrications

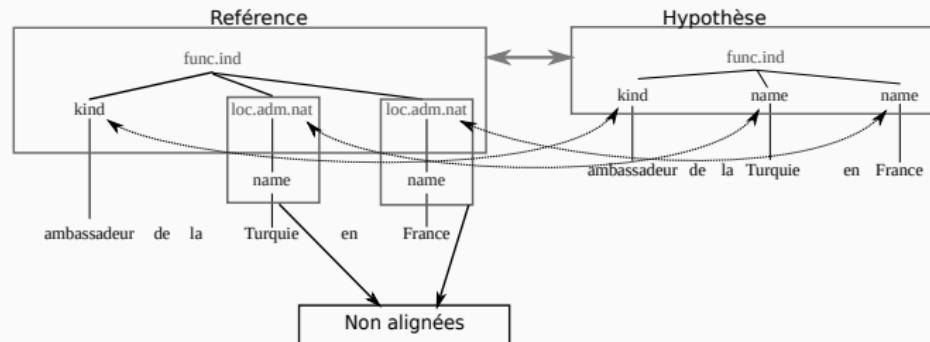
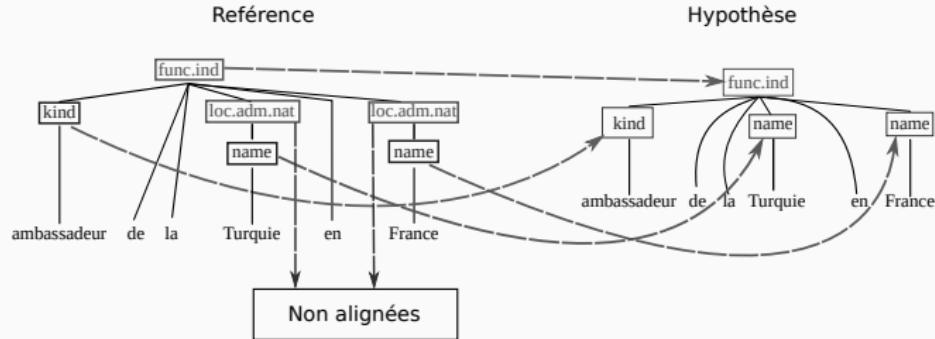
- représentation en “slot” des hypothèses et de la référence
 - slot= un segment de texte avec des frontières (début/fin) et un type
- structure plate qui ne peut pas traiter les entités imbriquées

ETER: Entity Tree Error Rate

Basée sur une comparaison des arbres d'entités [BJADG⁺14]

$$ETER = \frac{I + D + \sum_{(e_r, e_h)} E(e_r, e_h)}{N_E} \quad (10)$$

- I : nombre total d'insertions d'arbre-entité ;
- D : nombre total d'omission d'arbre-entité ;
- (e_r, e_h) : paires d'entités-arbres référence/hypothèse associées à l'issue de l'alignement ;
- $E(r, h)$: erreur calculée pour chaque paire d'entité-arbre (e_r, e_h) (peut être zéro) ;
- N_E : nombre d'entité-arbre dans la référence.



En haut, un

alignement basé sur les slots, en bas le même basé sur les arbres d'entités [BJ15]

$$ETER = \frac{I + D + \sum_{(e_r, e_h)} E(e_r, e_h)}{N_E} \quad (11)$$

Le calcul d'erreur pour les paires d'entité-arbre $E(r, h)$ a 2 parties

- erreur de détection et de classification de l'entité
- erreur de décomposition E_c

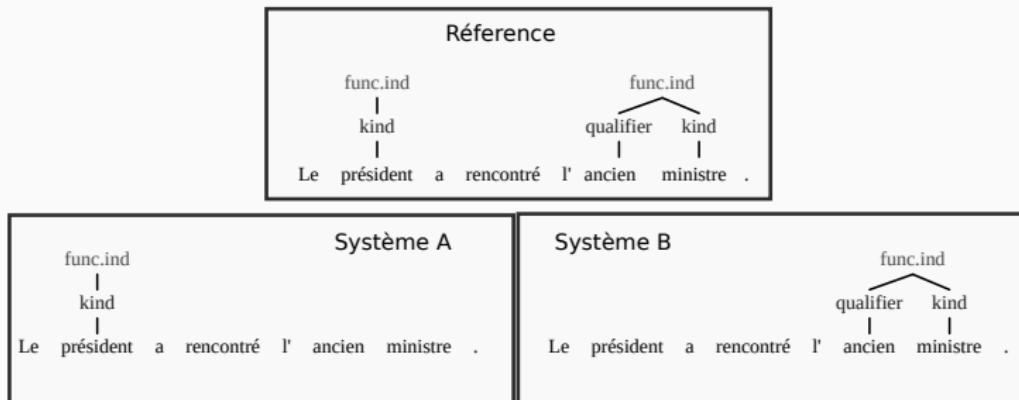
ETER: erreur de décomposition

$$E(r, h) = (1 - \alpha)E_T(e_r, e_h) + \alpha E_c(e_r, e_h), \alpha \in [0..1] \quad (12)$$

- $E_T(e_r, e_h)$: erreur de classification, dépend de la distance entre (e_r, e_h) ;
- $E_c(e_r, e_h)$: erreur de décomposition, dépend de la distance entre les constituants des entités-arbres (e_r, e_h) ;
- α paramètre fixant le poids relatif de la décomposition par rapport à la classification.

→ $E_c(e_r, e_h)$ se rapproche d'un SER local

ETER: exemple



- **Système A:** 3 omissions, 0 insertion, 0 substitution, 2 slots corrects
→ $3/5$ soit $SER = 60\%$
- **Système B:** 2 omissions, 0 insertion, 0 substitution, 3 slots corrects.
→ $2/5$ soit $SER = 40\%$

Or ces deux systèmes ont omis une entité chacun et devrait avoir un score équivalent. Avec *ETER*, chaque système présente 1 omission, 0 insertion et 0 erreur sur entité.

→ $ETER = 50\%$.

6. Evaluation

6.4 Evaluation des tâches connexes

Détection d'entités et de mentions

Détection et liaison

6. Evaluation

6.5 Evaluation des technologies appliquées en amont

Contacts

Maud Ehrmann
EPFL-DHLAB
maud.ehrmann@epfl.ch

Sophie Rosset
LIMSI
sophie.rosset@limsi.fr



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE



-  Mohamed Ameur Ben Jannet, *Évaluation adaptative des systèmes de transcription en contextes applicatifs*, Ph.D. thesis, Université Paris Sud, octobre 2015.
-  Mohamed Ameur Ben Jannet, Martine Adda-Decker, Olivier Galibert, Juliette Kahn, and Sophie Rosset, *Eter: a new metric for the evaluation of hierarchical named entity recognition*, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (Reykjavik, Iceland) (Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, eds.), European Language Resources Association (ELRA), may 2014 (english).

-  Razvan C Bunescu and Marius Pasca, *Using encyclopedic knowledge for named entity disambiguation.*, EACL, vol. 6, 2006, pp. 9–16.
-  UniProt Consortium et al., *The universal protein resource (uniprot) in 2010*, Nucleic acids research **38** (2010), no. suppl 1, D142–D148.
-  Nancy Chinchor, Ph.D and Beth Sundheim, *Muc-5 evaluation metrics*, Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993, 1993.
-  Silviu Cucerzan, *Large-scale named entity disambiguation based on wikipedia data.*, EMNLP-CoNLL, vol. 7, 2007, pp. 708–716.

-  George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel, *The automatic content extraction (ace) program, tasks, data, and evaluation*, Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004) (Lisbon, Portugal), European Language Resources Association (ELRA), May 2004, ACL Anthology Identifier: L04-1011.
-  Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin, *Entity disambiguation for knowledge base population*, Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 2010, pp. 277–285.
-  Ester2, *Entités nommées, dates, heures et montants*, 2007.

-  Nathalie Friburger, *Reconnaissance automatique des noms propres: application à la classification automatique de textes journalistiques*, Ph.D. thesis, Tours, 2002.
-  Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard, *Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview*, Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V) (Portland, OR), Association for Computational Linguistics, June 2011, pp. 92–100.
-  Ralph Grishman and Beth Sundheim, *Design of the muc-6 evaluation*, Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995, 1995.

References v

-  Heng Ji, Ralph Grishman, Hoa Trang Dang, and Joe Griffitt, Kira and Ellis, *Overview of the tac 2010 knowledge base population track*, Third Text Analysis Conference (TAC 2010), vol. 3, 2010, pp. 3–3.
-  J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, *Genia corpus: a semantically annotated corpus for bio-textmining*, no. suppl 1, i180–i182.
-  David McDonald, *Internal and external evidence in the identification and semantic categorization of proper names*, Corpus processing for lexical acquisition (1996), 21–39.
-  Paul McNamee and Hoa Trang Dang, *Overview of the tac 2009 knowledge base population track*, Text Analysis Conference (TAC), vol. 17, 2009, pp. 111–113.

-  Bethesda (MD), *UMLS® Reference Manual [Internet]*, Tech. report, National Library of Medicine (US), 09 2009.
-  John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel, *Performance measures for information extraction*, In Proceedings of DARPA Broadcast News Workshop, 1999, pp. 249–252.
-  David S. Pallett, *Performance Assessment of Automatic Speech Recognizers*, Res. National Bureau of Standards **90** (1985), no. 5, 371–387.
-  Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum, *Entités nommées structurées : guide d'annotation quaero*, LIMSI-CNRS, notes et documents limsi n° : 2011-04 ed., 2011.

-  D. Rao, P. McNamee, and M. Dredze, *Entity Linking: Finding Extracted Entities in a Knowledge Base*, Multi-source, Multilingual Information Extraction and Summarization, Springer, 2013, pp. 93–115.
-  Mihai Surdeanu and Heng Ji, *Overview of the english slot filling track at the tac2014 knowledge base population evaluation*, Proc. Text Analysis Conference (TAC2014), 2014.
-  Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela, *HAREM: An Advanced NER Evaluation Contest for Portuguese*, Irec (Genoa), May 2006, pp. 1640–1643.
-  Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata, *Extended named entity hierarchy*, LREC, 2002.

References viii

-  Beth M. Sundheim, *Overview of the third message understanding evaluation and conference*, THIRD MESSAGE UNDERSTANDING CONFERENCE (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991.
-  Erik F. Tjong Kim Sang and Fien De Meulder, *Introduction to the conll-2003 shared task: Language-independent named entity recognition*, Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (Stroudsburg, PA, USA), CONLL '03, Association for Computational Linguistics, 2003, pp. 142–147.
-  Vasudeva Varma, Praveen Bysani, Vijay Bharat Kranthi Reddy, Karuna Kumar Santosh GSK, Sudheer Kovelamudi, N Kiran Kumar, and Nitin Maganti, *iiit hyderabad at tac 2009*, Proceedings of Test Analysis Conference 2009 (TAC 09), 2009.