Corpus, ressources et linguistique outillée · M2SOL034

CM 5 : Reconnaissance des entités nommées (REN)

Ljudmila PETKOVIĆ

Semestre 2, 2024-2025
14 mars 2025

Sorbonne Université
Master « Langue et Informatique » (M1 ScLan)
UFR Sociologie et Informatique pour les Sciences Humaines

Cours adapté de Ehrmann et Rosset (2022).

Outline

LETTRES SORBONNE UNIVERSITÉ

Contexte et applications

Données non structurées

Une grande quantité de données (texte, images, audio. . .) est non structurée (sans modèle ni format pré-définis).

- comment exploiter les données, extraire l'information utile ?

Données	Information	Connaissance
description élémentaire d'une réalité	données avec un sens construisant une représentation de la réalité	informations avec une vérité
mesure des températures	courbe sur l'évolution des *minima* et *maxima* moyens en un lieu donné, par mois	fait que la température sur terre augmente du fait de l'activité humaine
série d'articles journalistiques	noms de personnes et leurs polarités	opinion des médias vis-à-vis de personnalités

Table 1 – Le qualitatif : données, informations et connaissances.

LETTRES SORBONNE UNIVERSITÉ

L'information est « cachée » dans les textes

Extraction d'information (EI) : extraire des informations structurées à partir de textes non structurés :

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

Exemple :

On the invitation of the Festival de Cannes, the Italian actress Monica Bellucci 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

LETTRES SORBONNE UNIVERSITÉ

Principales tâches en EI

- **traitement des entités nommées (EN)**
  reconnaissance, catégorisation et désambiguïsation
  - *Monica Bellucci* et *Pedro Almodovar* → PERSON
  - *Monica Bellucci* $\xrightarrow{\text{ref.}}$ https://dbpedia.org/page/Monica_Bellucci
- **traitement des expressions temporelles**
  extraction et normalisation
  - *from 17 to 28 May 2017* → DURATION
  - *from 17 to 28 May 2017* → [17-05-2017, 28-05-2017]
- **extraction d'évènements**
  - *70th Festival de Cannes* → FACTUAL, RECURRING EVENT
  - *70th Festival de Cannes* → $\xrightarrow{\text{instance\_of}}$ https://fr.wikipedia.org/wiki/Festival_de_Cannes
- **extraction de relations**
  - *70th Festival de Cannes*, TOOKPLACE, [17-05-2017, 28-05-2017]

LETTRES SORBONNE UNIVERSITÉ

Définition

Entités nommées : définition et tâches

'90 : campagnes d'évaluations sur la compréhension de documents

- éléments d'intérêt : PERSONNE, ORGANISATION, LIEU etc.
- unités référentielles qui sous-tendent la sémantique des textes

1. reconnaissance : détecter les EN 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'EN à une référence unique (on lie à une référence)
4. extraction de relation : découvrir des relations entre entités (FATHER-OF, BORN-IN, ALMA MATER)

LETTRES SORBONNE UNIVERSITÉ

Les EN dans le monde : le problème de la catégorisation

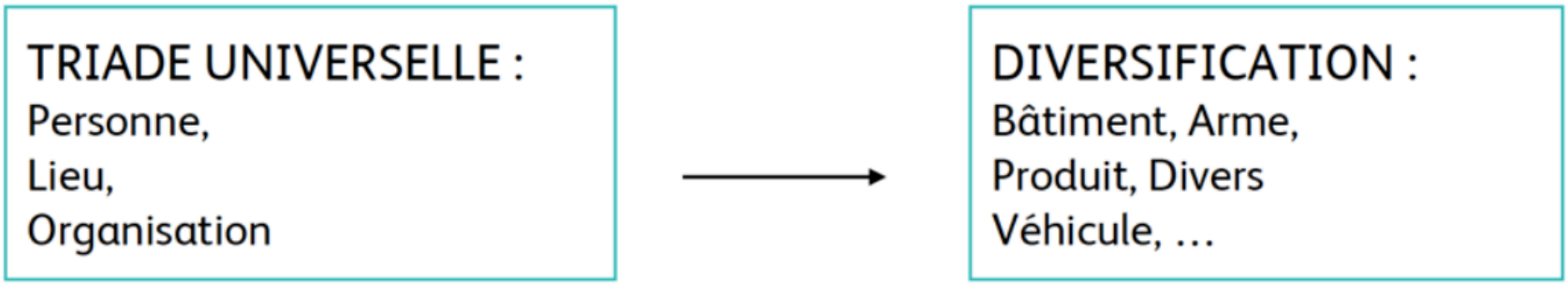


Figure 1 – Le choix des catégories des EN.

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	...

Figure 2 – La détermination de ce que les EN recouvrent.

→ catégorisation instable

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

- **Combinaisions de syntagmes : une ou plusieurs entités ?**
  - *Les banques centrales américaine et européenne ont décidé...*
  - *Donald et Melania Trump*
  - *l'université de Genève*
- **Un syntagme : quelles frontières ?**
  - *la candidate Ségolène Royal, Professeur Paolucci*
  - *George W. Bush Jr., La Mecque, l'Abbé Pierre*
- **Une entité : quelle unité lexicale ?**
  - *Émmanuel Macron*, *Monsieur Macron*, *le Président Émmanuel Macron*, *le Président français*, *le Président de la République française*, *Manu*

→ caractérisation imprécise, diversité des mentions

LETTRES SORBONNE UNIVERSITÉ

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'État, a changé de position sur la présence française au sein de la force internationale.*

→ polyréférentialité

EN : un objet TAL difficile à cerner

- **Hétérogénéité des réalisations**
  - Les EN ne se limitent pas à une catégorisation, une mention, une interprétation
- **Hétérogénéité des points de vue**
  - formules définitoires sous la forme d'énumérations
  - caractérisation diverses (sens, forme)

→ Question : que sont les EN ?

Le « matériau » de départ

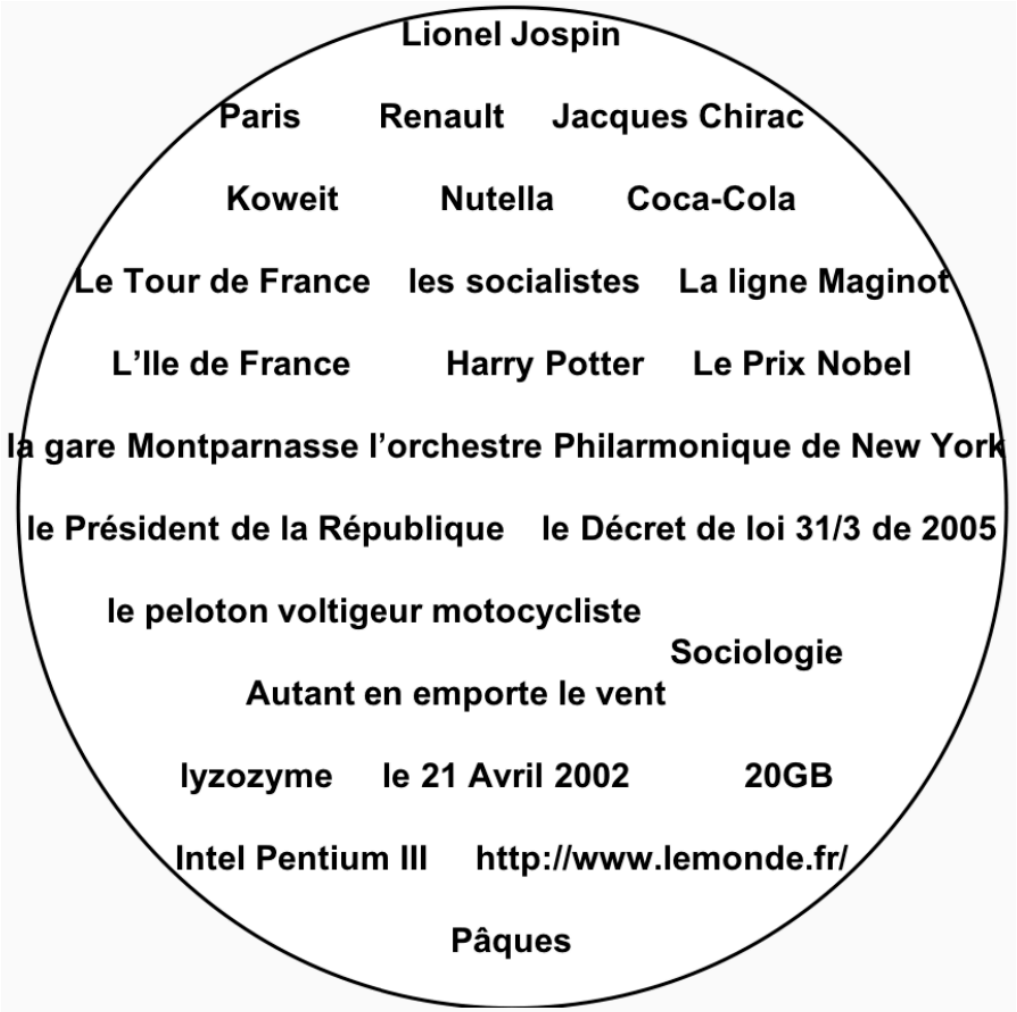


Figure 3 – Unités lexicales considérées comme des EN.

Le « matériau de départ »

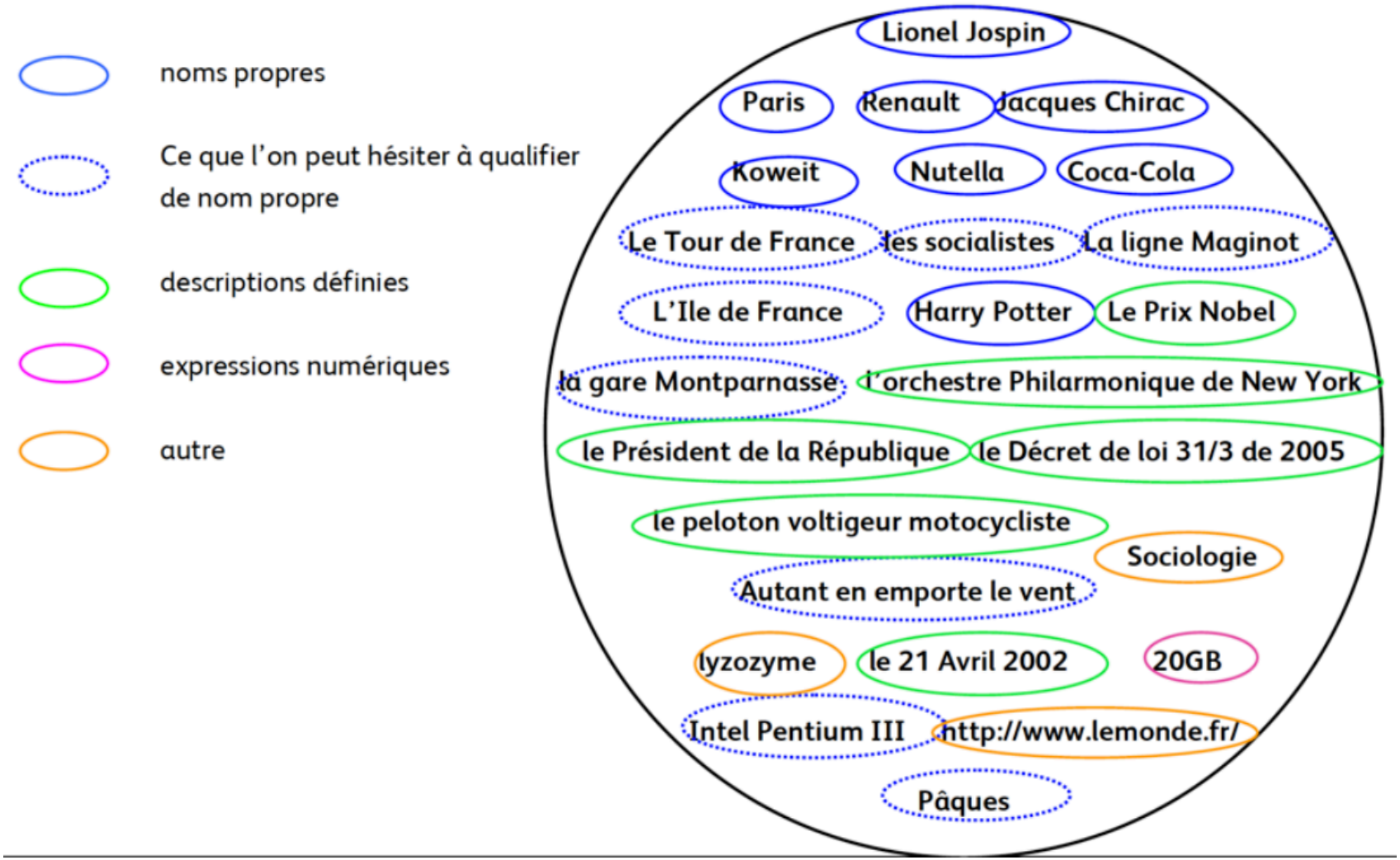


Figure 4 – Proposition de catégorisation des EN.

Unicité référentielle

- **Le nom propre se 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'existants
  - **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

LETTRES SORBONNE UNIVERSITÉ

Autonomie référentielle

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 (in)complètes
  - *le président*, *le président de la République française en 2003*

Caractérisation linguistique des EN

- L'ensemble EN 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

Proposition de définition

Entité nommée

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

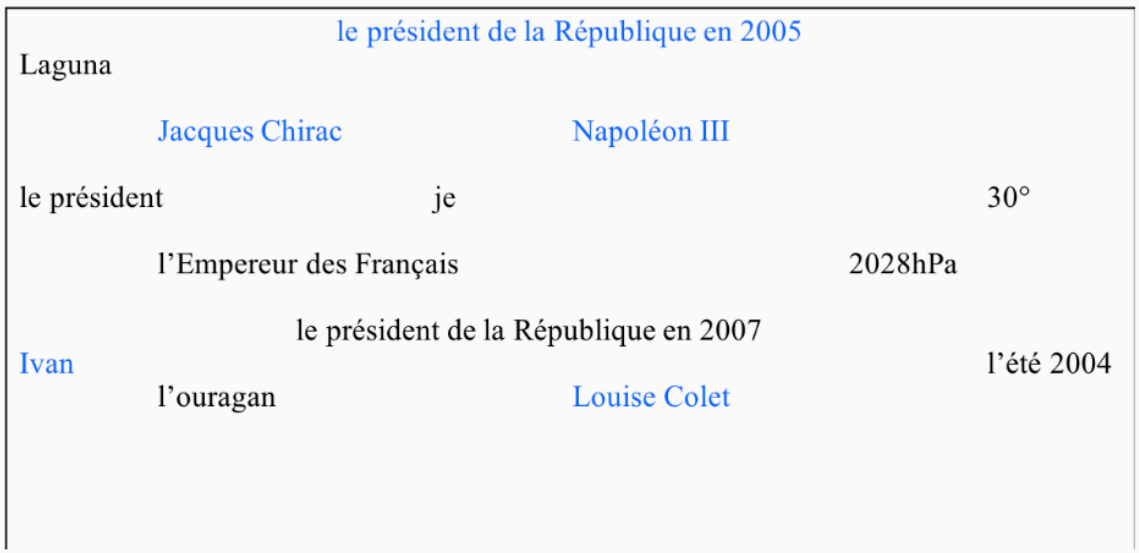
Questions que l'on s'est posées :

- 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 ?

LETTRES SORBONNE UNIVERSITÉ

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.

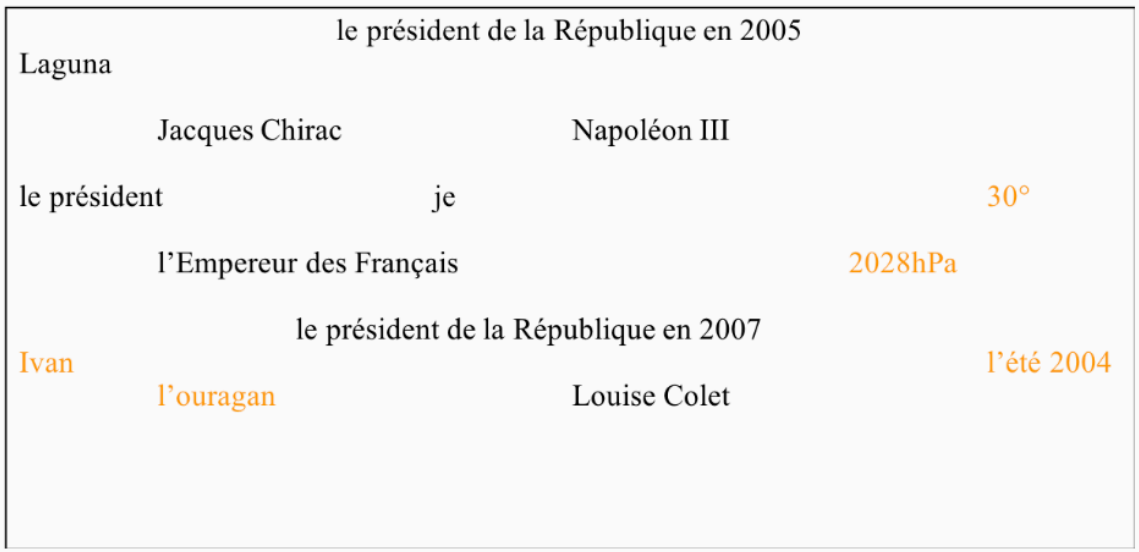


Application : générique « typique »
Modèle : Personnes, Lieux, Organisations
Corpus : journalistique français de 1998 à 2008

Figure 5 – Cas de figure I.

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.

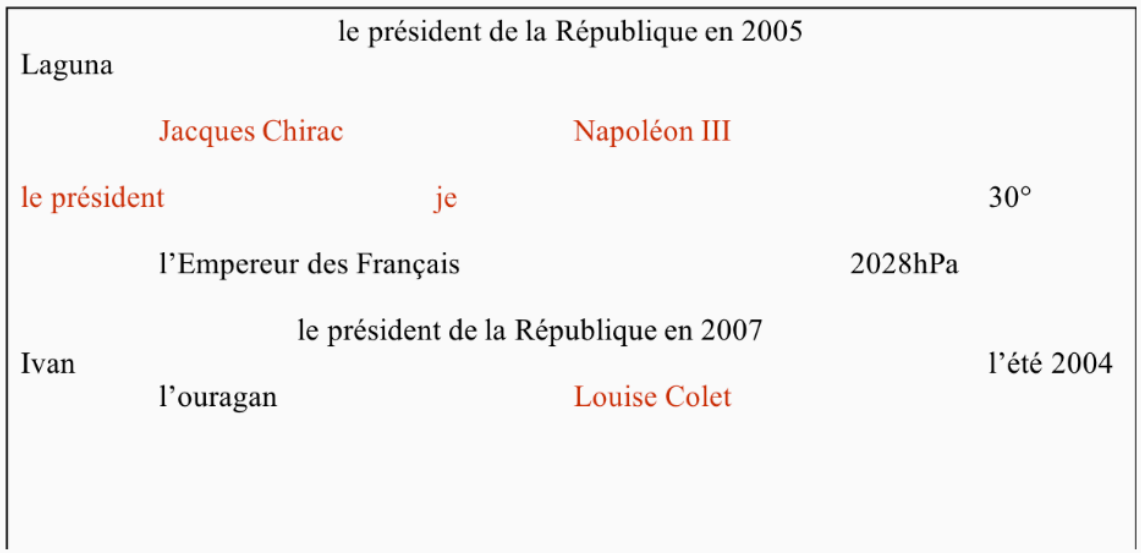


Application : étude sur le climat
Modèle : températures, mesures atmosphérique, ouragan, dates, périodes, …
Corpus : totalité des observations météorologiques sur une période données

Figure 6 – Cas de figure II.

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.



Application : « littéraire »
Modèle : personnes, lieux, événements
Corpus : correspondance de Flaubert

Figure 7 – Cas de figure III.

Les EN : une création TAL

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

$\rightarrow$ pas d'EN *per se*, seulement des critères linguistiques et un modèle

Conséquences : points de vue

- général : explication de l'hétérogénéité et de la variabilité de l'ensemble EN
- pratique : critères de décision pour annoter

LETTRES SORBONNE UNIVERSITÉ

Ressources

Ressources

De quoi a-t-on besoin pour traiter les EN ?

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

LETTRES SORBONNE UNIVERSITÉ

Typologie : une façon de structurer

Une typologie (angl. *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
- structure
- sur la définition de ce que recouvrent les catégories

LETTRES SORBONNE UNIVERSITÉ

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
DATE	en **juillet**	en **juillet** dernier

Table 2 – Le qualitatif : données, informations et connaissances.

LETTRES SORBONNE UNIVERSITÉ

Typologie ACE

Types	Sous-types
PERS	individu, groupe, indéterminé
ORG	(non) gouvernementales, commerciales, éducation, divertissement, média, religieuses, médicales, 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, 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 spécifiées

Évolution

Nombreuses autres typologies s'inspirant de MUC et ACE

- CoNLL : inspiration MUC, ajout d'une catégorie MISC
- HAREM : inspiration ACE, ajout de différentes catégories (Idée, Objet, Autre, Groupe)
- ESTER-2 : encore plus de sous-types (PERS.HUM, PERS.ANIM, LOC.GEO, LOC.ADMIN etc.) et traitement de l'imbrication

Imbrication des EN

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

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

Structuration très utilisée dans des domaines de spécialité, p. ex. la typologie GENIA (domaine bio-médical)

La typologie QUAERO

1. **Personne** : personne individuelle, groupe de personnes
2. **Lieu** : lieu administratif, lieu physique, construction, odonyme, adresse
3. **Organisation** : administration, service
4. **Expression temporelle** : date / heure absolue et relative
5. **Montant**
6. **Produit** : objet manufacturé, route, produit financier, doctrine, loi, *software*, art, média, récompense
7. **Fonction** : individuelle ou collective

Typologie QUAERO : sous-types

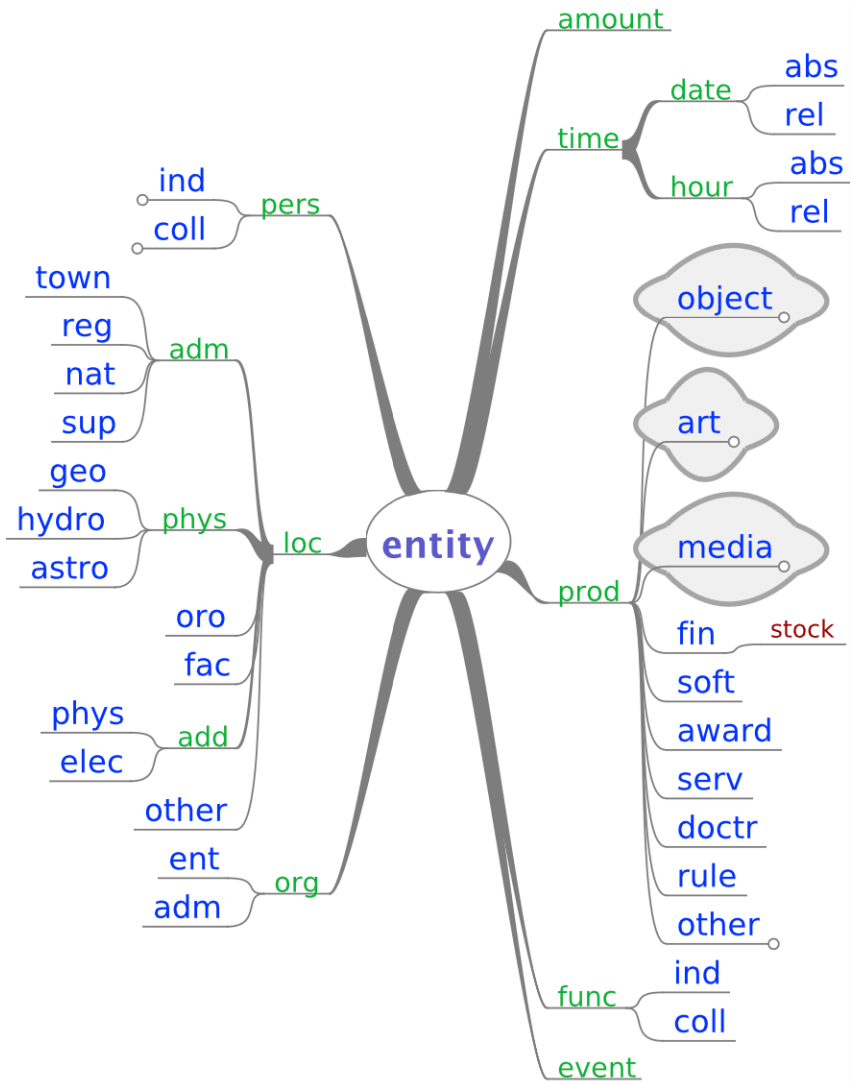


Figure 8 – Les sous-types de la typologie QUAERO.

Typologie QUAERO : composants d'entités

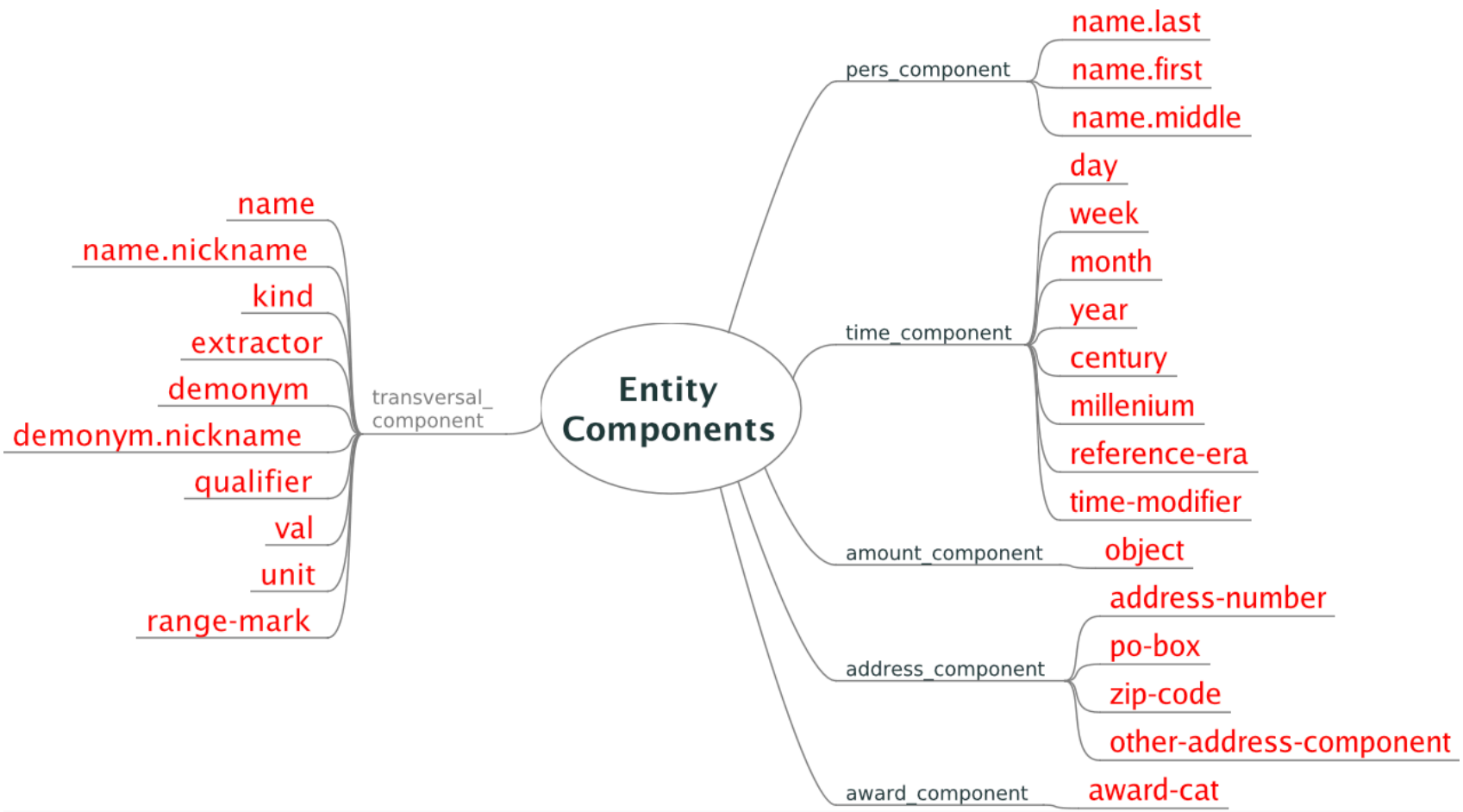


Figure 9 – Les composants d'entités de la typologie QUAERO.

QUAERO : composants d'entités

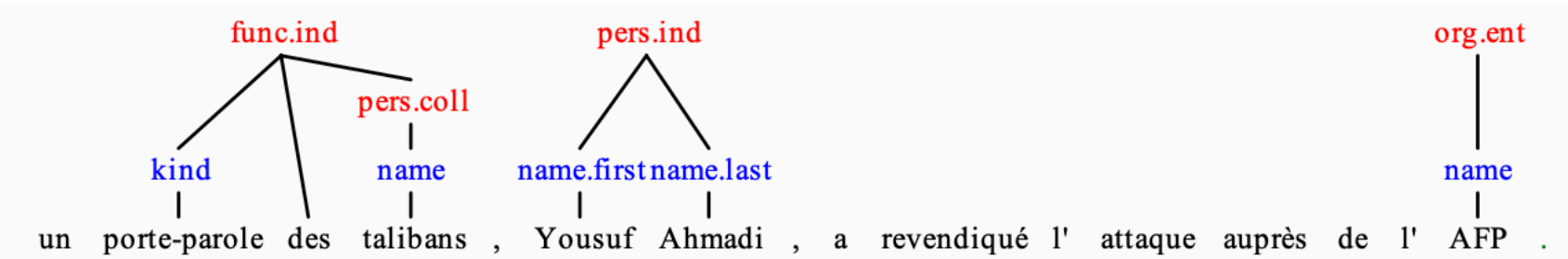


Figure 10 – Les composants d'entités de la typologie QUAERO.

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

Table 4 – Comparaison de typologies des EN.

LETTRES SORBONNE UNIVERSITÉ

Text Analysis Conference – Knowledge Base Population

Pour une EN donnée, il importe de trouver de nombreux attributs.
P. ex. pour une entité de type PERS :

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

LETTRES SORBONNE UNIVERSITÉ

Corpus annoté et guide d'annotation

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

Typologie → 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

LETTRES SORBONNE UNIVERSITÉ

Campagne d'annotation

- à partir d'outils dédiés (BRAT [1], GLOZZ [2], WEBANNO [3])
- 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 : chronophage et gourmand en ressources

Exemples de corpus français : ESTER 2, QUAERO, ETAPE

1. https://brat.nlplab.org/introduction.html
2. http://explorationdecorpus.corpusecrits.huma-num.fr/glozz/
3. https://webanno.github.io/webanno/

LETTRES SORBONNE UNIVERSITÉ

Lexiques et bases de connaissances

Objectif : fournir des informations relatives à des EN, 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.

Types d'informations :

1. lexicales, sur les unités composant les EN
2. encyclopédiques, sur les référents des EN

Évolution importante de ce type de ressource depuis l'apparition de la tâche : **index géographiques** (angl. *gazetteers*) → encodage de plus en plus d'information

LETTRES SORBONNE UNIVERSITÉ

Bases lexicales

Encodent 2 types d'information :

- des **noms ou parties de noms d'entités** avec leurs types associés, p. ex. *Justin* $\rightarrow$ directement utilisés pour reconnaître des unités équivalentes dans les textes
- des **mots amorces**, également avec leurs types associés, p. ex. *Monsieur* $\rightarrow$ des unités indiquant avec une forte probabilité la présence d'une EN d'un certain typye

- WordNet [4] : utile pour l'intégration de ressources
- PROLEX [5] : base d'EN multilingue
- Geonames [6] : toponymes et assimilés

4. https://wordnet.princeton.edu/
5. https://www.ortolang.fr/market/lexicons/prolex
6. https://www.geonames.org/

LETTRES SORBONNE UNIVERSITÉ

Reconnaissance et classification

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'EN
- **robustesse** : ne pas échouer face à des cas non canoniques

En pratique :

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

LETTRES SORBONNE UNIVERSITÉ

Représentation du texte

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

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

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

- des caractères : **indices morphologiques**
  - majuscule, régularités socio-culturelles (-*ville*), nombres
- des mots eux-mêmes : **indices lexicaux**
  - confronter les textes à des listes d'EN de composants d'EN
- de la séquence de mots : **indices contextuels**
  - contextes local (mots qui précèdent ou suivent l'EN) et global (phrase(s), etc.)

LETTRES SORBONNE UNIVERSITÉ

Approches symboliques

Techniques à base d'automates

- insertion de balises dans les textes indiquant où se trouvent les EN
- conception de **règles** formant une **grammaire locale**
- boîtes à outils : Unitex, GATE, NooJ, etc.

Pré-traitements : segmentation en mots, en phrases, étiquetage morphosyntaxique

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

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'atout pour le contrôle et de l'ingénierie : plus compréhensibles, modulables, possibilités de réglages fins
- majoritaires dans le milieu industriel

LETTRES SORBONNE UNIVERSITÉ

Apprentissage automatique

Modèles **guidés par les données** (angl. *data-driven* models)

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 sur les données.

LETTRES SORBONNE UNIVERSITÉ

Le paradigme de l'apprentissage automatique

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

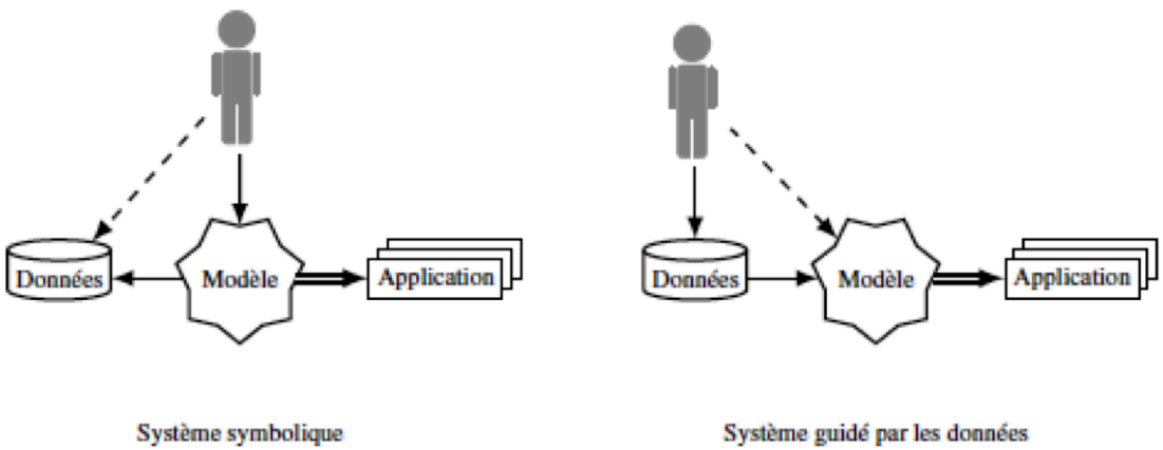


Figure 11 – Système symbolique *vs.* système guidé par les données.

Approches existantes

La REN peut être formalisée comme une tâche de **classification**.

- arbres de décision
- modèles probabilistes
- réseaux neuronaux

Modèles par classes majoritaires

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 des probabilités :

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

LETTRES SORBONNE UNIVERSITÉ

Modèles par classes majoritaires

La formule de Bayes et l'estimation statistique permettent de calculer la probabilité d'une étiquette étant donné 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 étiquette pour un mot donné = ratio entre la fréquence du mot avec une étiquette dans le corpus annoté et la fréquence du mot dans ce même corpus, quelle que soit l'étiquette.

Modèles par classes majoritaires

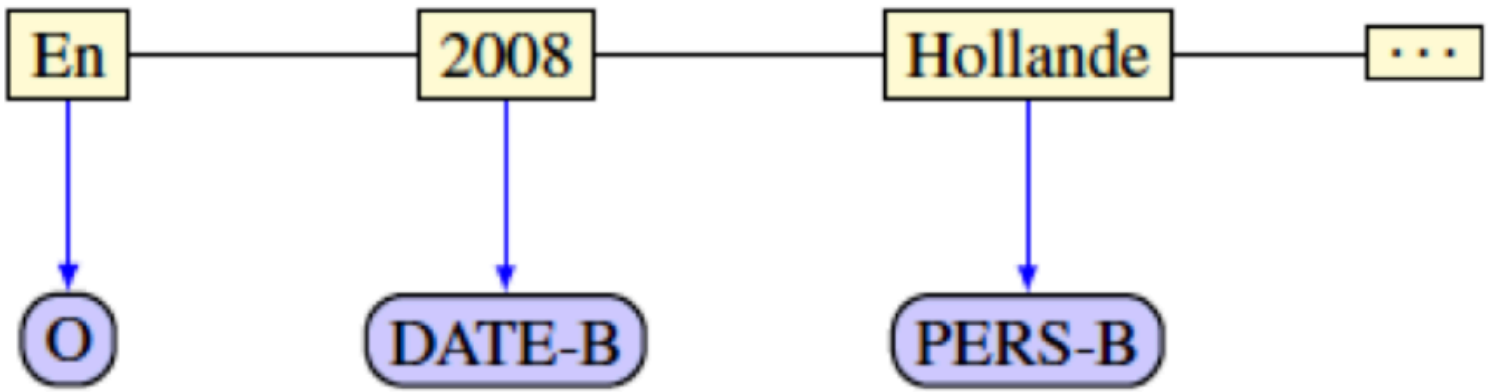


Figure 12 – Modèle par classes majoritaires. L'orientation des flèches indique quelles dépendances sont prises en compte par le modèle.

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 inverse : déterminer, pour une suite d'étiquettes, la probabilité qu'elle génère un texte donné.

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

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

LETTRES SORBONNE UNIVERSITÉ

Modèles à décisions contextuelles (HMM)

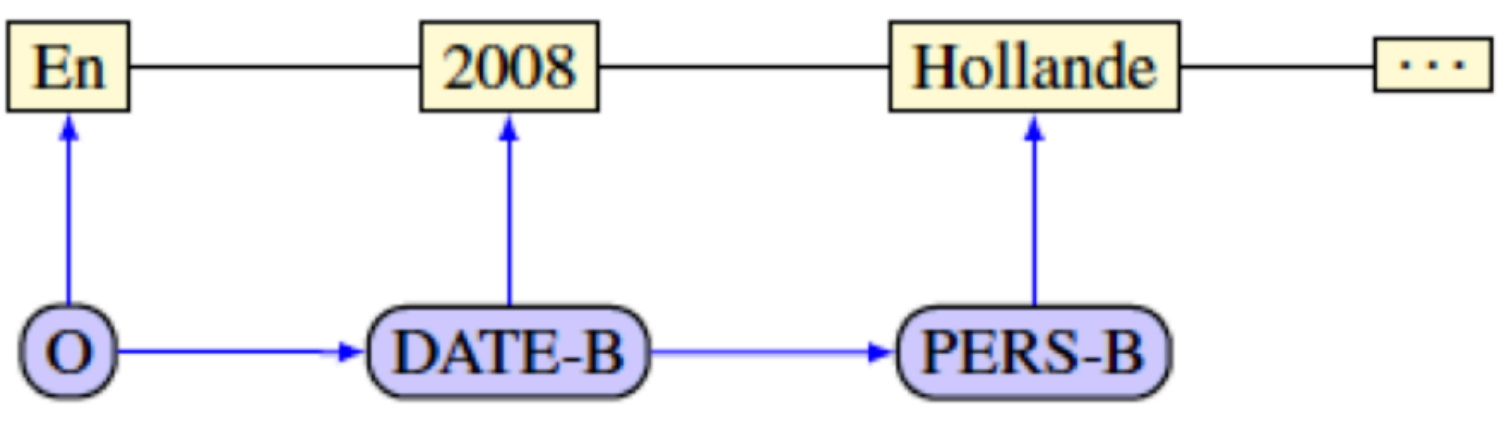


Figure 13 – Modèle de Markov à états cachés. Décisions non indépendantes : la solution la plus vraisemblable est choisie en fonction des étiquettes préalablement choisies.

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.

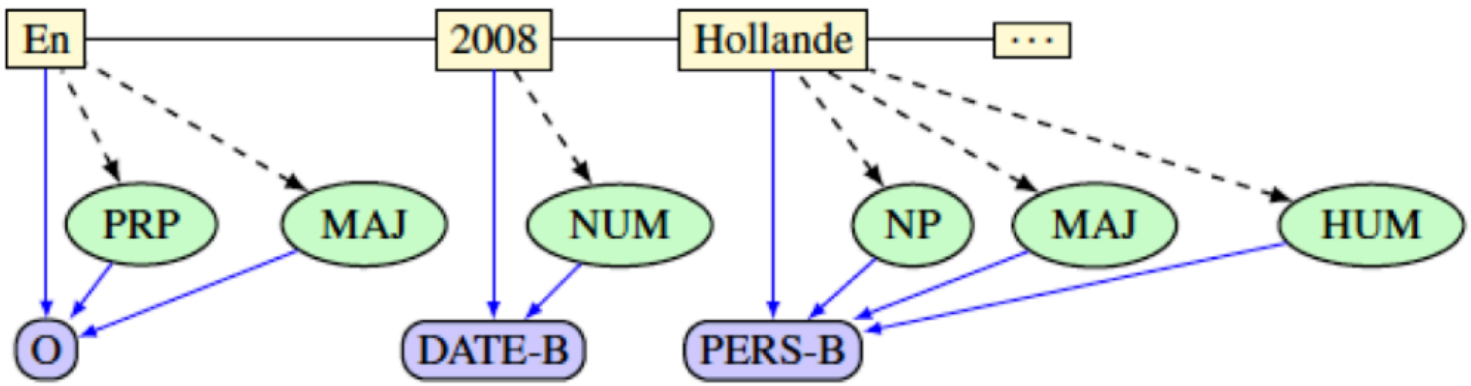


Figure 14 – Tenir compte d'indice sur les tokens.

Champs aléatoires conditionnels (CRF)

Les **CRF** (angl. *Conditional Random Fields*) ou champs aléatoires 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 pré-traitements)

Champs aléatoires conditionnels (CRF)

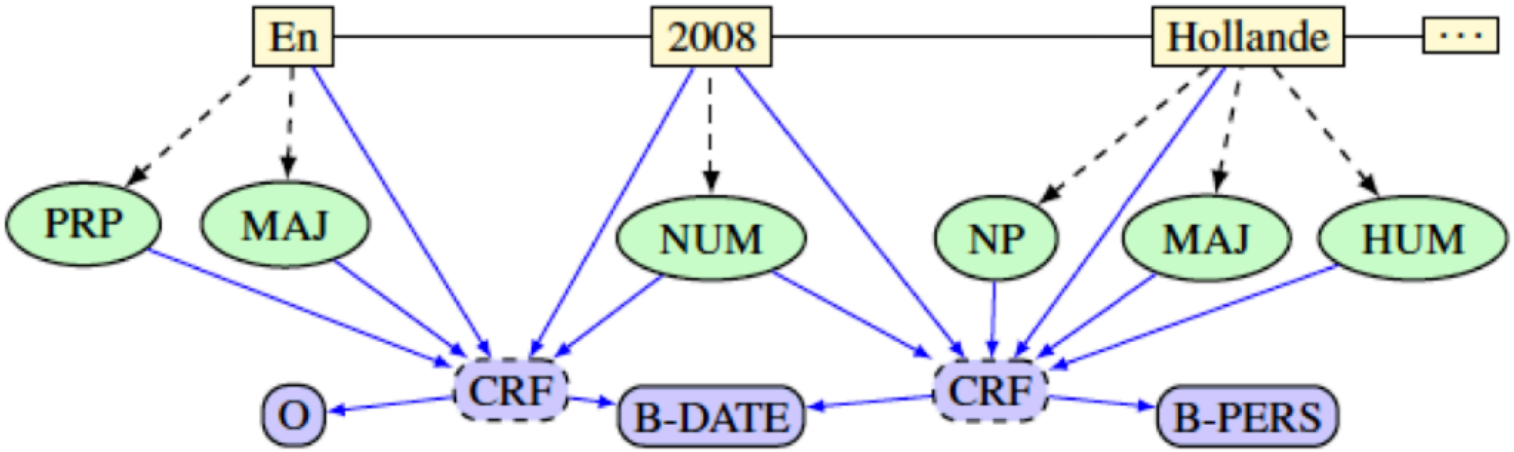


Figure 15 – Modèle graphique CRF.

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

Fonction exponentielle pour évaluer la pertinence d'un état donné en fonction d'un ensemble de caractéristiques.

Liaison

Où en sommes-nous ?

- nous savons reconnaître et catégoriser des segments textuels : des **mentions** d'EN 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 se réfèrent

→ objectif : **désambiguïsation, résolution, liaison**

Des mentions aux référents

Catégoriser n'est pas désambiguïser

- *G. Bush* et *F. Mitterrand* sont des PERSON
- lequel des deux se réfère-t-il au 43e président des États-Unis ?

Le problème des homonymes

- *F. Mitterrand* est une PERSON (*François* ou *Frédéric* ?)
- *Bush* est une PERSON (*G. Bush* ou *G. W. Bush* ?)

Le problème des variantes

- *Jean-Claude Juncker*, *Juncker* et *le président de la Commission Européenne* se réfèrent-elles à la même EN ?

LETTRES SORBONNE UNIVERSITÉ

Le point sur les tâches

- **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 donné des documents, identifier les mentions d'EN et lier chacune d'elles à un référent d'une base de connaissances

LETTRES SORBONNE UNIVERSITÉ

Évaluation

Évaluer

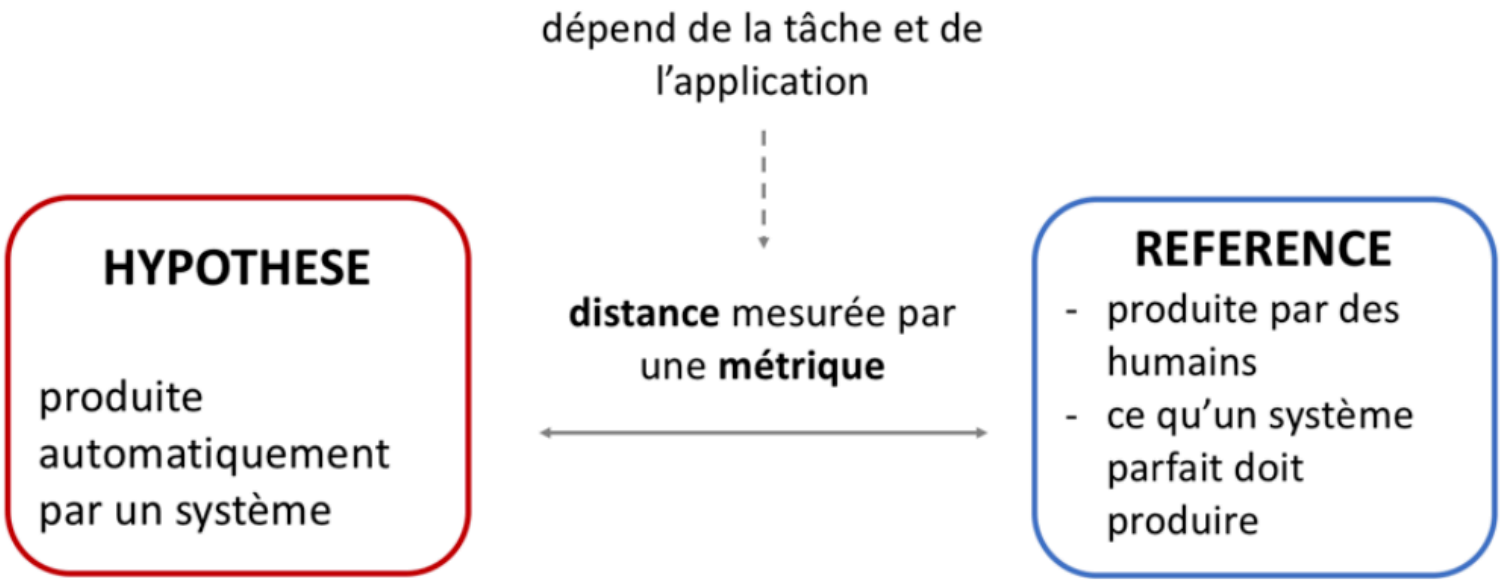


Figure 16 – Objectif : mesurer à quel point le système trouve les « bonnes réponses ».

Quelles « bonnes réponses » ?

- traduction ou résumé automatique : bonnes réponses multiples
- REN : une seule et unique bonne réponse

Avantages et exigences

- **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
3. un **algorithme de projection** des EN annotées sur la transcription manuelle de référence vers la transcription automatique

LETTRES SORBONNE UNIVERSITÉ

Les mesures classiques

Précision

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

$$P = \frac{C}{C + S + I}$$

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

LETTRES SORBONNE UNIVERSITÉ

Les mesures classiques

Rappel

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

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

- D : nombre total d'**omissions** (suppressions) opérées par le systèmes (EN non détectées, silence)
- $C + S + D$: 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>

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

$\rightarrow$ 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

- Précision : $\frac{2}{2} = 1$
- Rappel : $\frac{1}{2} = 0,5$

$\rightarrow$ HYP2 produit du **silence**

F-mesure

Définie comme la moyenne harmonique entre Précision et Rappel :

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

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

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

$$F(HYP2) = (1 + 1^2) \times \frac{1 \times 0,5}{1^2 \times 1 + 0.5} = 0,67$$

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
- avec les typologies fines et complexes, besoin d'une métrique différenciant les erreurs

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

HYP1 : `the <pers.ind>president</pers.ind> of Ford`

$\rightarrow$ erreur de frontière

HYP2 : `the <pers.coll>president of Ford</pers.coll>`

$\rightarrow$ erreur de sous-type

HYP3 : `the <pers.coll>president</pers.coll> of Ford`

$\rightarrow$ erreur de sous-type et de frontière

LETTRES SORBONNE UNIVERSITÉ

Mesures basées sur le décompte d'erreurs : SER

SER : *Slot Error Rate* (Makhoul et al., 1999)

- identique au *WER* utilisé en reconnaissance autom. de parole
- utilisée lors de ACE, ESTER-2, QUAERO et ETAPE
- suppression du nombre d'insertion (I) du dénominateur

$$SER = \frac{S + D + I}{C + D + S} = \frac{S + D + I}{R}$$

où $R =$ nombre total d'EN de la référence

SER

Possibilité d'affiner l'importance relative des erreurs

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

- S_t et S_f : nb total de substitutions de type et de frontières
- D et I : nombre total d'omissions et d'insertions
- α_1, α_2, β et γ : poids affectées à chaque catégories d'erreur

Références


Ehrmann, M. et S. Rosset (2022). ***Le Traitement des Entités Nommées · définition, ressources, méthodes, applications.*** Diapositives, https://github.com/BigDataSpeech/EN/blob/gh-pages/docs/classEN.pdf. Consulté le 14 mars 2025 (*voir p. 1*).


Makhoul, J., F. Kubala, R. Schwartz, R. Weischedel et al. (1999). **Performance measures for information extraction.** In : *Proceedings of DARPA broadcast news workshop*. http://ccc.inaoep.mx/~villasen/bib/slot%20error%20rate.pdf. Herndon, VA, p. 249-252 (*voir p. 71*).

Licence