Désambiguïsations lexicales
Langues et entités nommées

Damien Nouvel

ERTIM
Equipe de recherche
textes, informatique,
multilinguisme

inalco
Institut national
des langues
et civilisations orientales

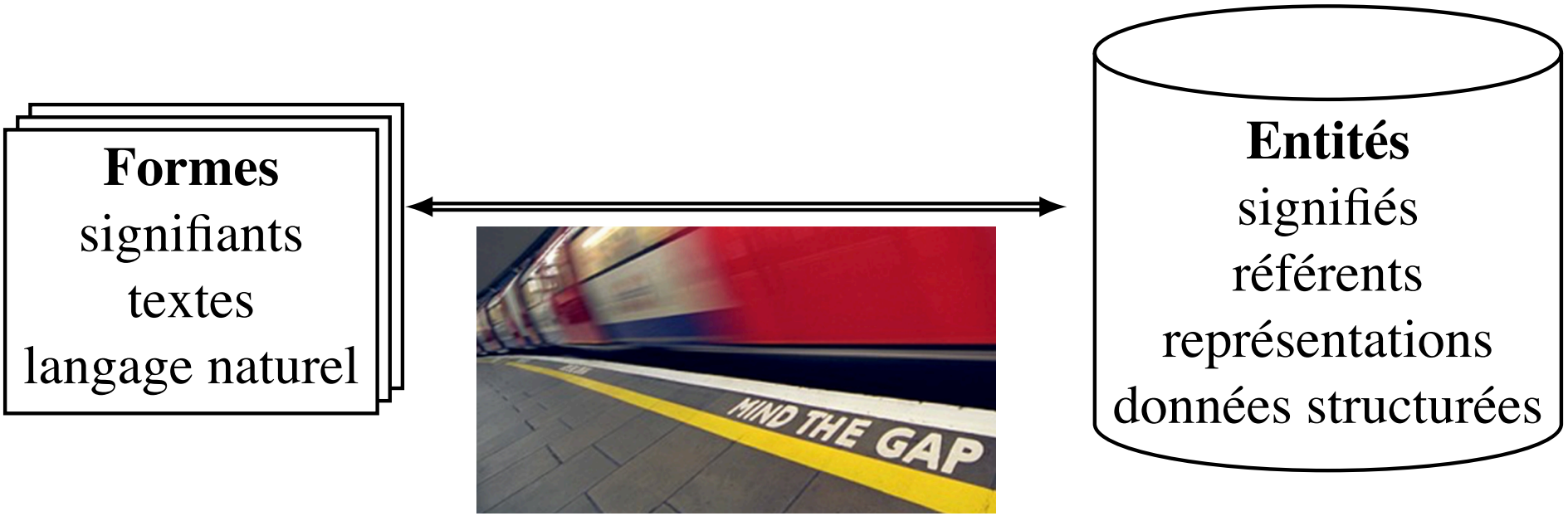
http://damien.nouvels.net/bazar/desamb.pdf

Plan

http://damien.nouvels.net/bazar/desamb.pdf

Des formes aux sens …

Des formes aux sens …



- Comment faire le **lien** entre **textes** et **bases de connaissances** ?
- Quel **niveau** d'analyse : caractères, mots, énoncés, documents ?
- Comment représenter le **contexte** ?

$\Rightarrow$ Processus de **compréhension** (vs génération)

De l'ambiguïté

« MAGNIFIQUE PORTE ! »

$\Rightarrow$ Ce texte ne semble guère ambigu …

De l'ambiguïté

« MAGNIFIQUE PORTE ! »

$\Rightarrow$ Ce texte ne semble guère ambigu …

« Magnifique porté ! »

$\Rightarrow$ Désambiguisation par le lexique (système d'écriture)

De l'ambiguïté

« MAGNIFIQUE PORTE ! »

⇒ Ce texte ne semble guère ambigu …

« Magnifique porté ! »

⇒ Désambiguisation par le lexique (système d'écriture)

« CE PATINEUR A FAIT UN MAGNIFIQUE PORTE ! »

⇒ Désambiguisation contextuelle

Désambiguisation : situations et enjeux

- La désambiguisation comme **compréhension**

Désambiguisation : situations et enjeux

- La désambiguisation comme **compréhension**
- Sans communication
  - Sens et représentation dans le **cerveau** (imagerie)
  - Représentations cognitives / **neuronales**

  $\Rightarrow$ Quelle existence pour le langage sans communication ?

Désambiguisation : situations et enjeux

- La désambiguisation comme **compréhension**
- Sans communication
  - Sens et représentation dans le **cerveau** (imagerie)
  - Représentations cognitives / **neuronales**

  $\Rightarrow$ Quelle existence pour le langage sans communication ?
- Communication interpersonnelle (« proches »)
  - Dialogue en interaction : peu d'ambiguïtés

  $\Rightarrow$ Aspect visuel (gestes, visage, etc.)

  $\Rightarrow$ Contextualisation des énoncés

  $\Rightarrow$ À l'écrit : des langages **prescrits** vs **décrits**

Désambiguisation : situations et enjeux

- La désambiguisation comme **compréhension**
- Sans communication
  - Sens et représentation dans le **cerveau** (imagerie)
  - Représentations cognitives / **neuronales**

  $\Rightarrow$ Quelle existence pour le langage sans communication ?
- Communication interpersonnelle (« proches »)
  - Dialogue en interaction : peu d'ambiguïtés

  $\Rightarrow$ Aspect visuel (gestes, visage, etc.)

  $\Rightarrow$ Contextualisation des énoncés

  $\Rightarrow$ À l'écrit : des langages **prescrits** vs **décrits**
- Variations de l'intercommunication
  - Prononciations, dialectes, bruit
  - Le langage comme **convention sociale**
  - Langues de **spécialité** (domaines)
  - Grande diversité des **systèmes d'écritures** (150)

Cadre général

- Modalités
  - Oral / spontané
  - Écrit
  - Signé

Cadre général

- Modalités
  - Oral / spontané
  - Écrit
  - Signé
- Niveaux d'analyse
  - Morphèmes / phonèmes
  - Mots / lexèmes
  - Syntagmes
  - Propositions
  - Unités discursives
  - Documents

Cadre général

- Modalités
  - Oral / spontané
  - Écrit
  - Signé
- Niveaux d'analyse
  - Morphèmes / phonèmes
  - Mots / lexèmes
  - Syntagmes
  - Propositions
  - Unités discursives
  - Documents
- Autres informations **contextuelles** (auteur, date, etc.)

Cadre général

- Modalités
  - Oral / spontané
  - Écrit
  - Signé
- Niveaux d'analyse
  - Morphèmes / phonèmes
  - Mots / lexèmes
  - Syntagmes
  - Propositions
  - Unités discursives
  - Documents
- Autres informations **contextuelles** (auteur, date, etc.)
- Systèmes actuels : majoritairement **propositions** / phrases

Multilinguisme

- Systèmes d'écritures
  - Nécessité de **segmenter** (chinois, thaï)
  - Absence de **diacritiques** : arabe, hébreu, etc.
  - Avec ou sans casse (noms propres)
  - Termes empruntés / translittérés

Multilinguisme

- Systèmes d'écritures
  - Nécessité de **segmenter** (chinois, thaï)
  - Absence de **diacritiques** : arabe, hébreu, etc.
  - Avec ou sans casse (noms propres)
  - Termes empruntés / translittérés
- Compréhension qui s'appuie sur
  - Les **lexiques**
  - La **morphologie**
  - La **syntaxe**

  $\Rightarrow$ À exploiter dans des systèmes

Multilinguisme

- Systèmes d'écritures
  - Nécessité de **segmenter** (chinois, thaï)
  - Absence de **diacritiques** : arabe, hébreu, etc.
  - Avec ou sans casse (noms propres)
  - Termes empruntés / translittérés
- Compréhension qui s'appuie sur
  - Les **lexiques**
  - La **morphologie**
  - La **syntaxe**

  $\Rightarrow$ À exploiter dans des systèmes

$\Rightarrow$ Disparité des ressources lexicales selon les langues

$\Rightarrow$ Difficile de faire des systèmes uniformes …

La translittération

- Texte sans voyelles :

نقوش فصوص خواتيم الفلاسفة

Nqwsh Fsws Khwatym Al-Flasft

- Texte avec voyelles

نُقُوشِ فَصُوصِ خَواتِيمِ الفَلاسِفَةِ

Nuqoushin Fasousi Khawateimi Al-Falasifati

⇒ Importante pour les (nouveaux) noms propres !

Plan

http://damien.nouvels.net/bazar/desamb.pdf

Entités nommées

Exemples d'entités nommées

'Rappeur, chanteur et compositeur **congolais** *né le* **6 mai 1986** *à* **Kinshasa**. *Il est issu d'une famille de musiciens, son père était un chanteur du groupe* **Viva La Musica** *de* **Papa Wemba**.*'*

Entités nommées

- Extraction d'informations au sein du langage naturel

Exemples d'entités nommées

'Rappeur, chanteur et compositeur **congolais** *né le* **6 mai 1986** *à* **Kinshasa**. *Il est issu d'une famille de musiciens, son père était un chanteur du groupe* **Viva La Musica** *de* **Papa Wemba**.'

Entités nommées

- Extraction d'informations au sein du langage naturel
- Deux types d'expressions linguistiques concernées
  - **Noms propres** : personnes, lieux, organisations, produits
  - **Descriptions définies** : expressions de temps, montants, fonctions

Exemples d'entités nommées

'Rappeur, chanteur et compositeur **congolais** *né le* **6 mai 1986** *à* **Kinshasa**. *Il est issu d'une famille de musiciens, son père était un chanteur du groupe* **Viva La Musica** *de* **Papa Wemba**.'

Entités nommées

- Extraction d'informations au sein du langage naturel
- Deux types d'expressions linguistiques concernées
  - **Noms propres** : personnes, lieux, organisations, produits
  - **Descriptions définies** : expressions de temps, montants, fonctions

Exemples d'entités nommées

'Rappeur, chanteur et compositeur **congolais** *né le* **6 mai 1986** *à* **Kinshasa**. *Il est issu d'une famille de musiciens, son père était un chanteur du groupe* **Viva La Musica** *de* **Papa Wemba**.*'*

$\Rightarrow$ Comment « trouver » les entités nommées dans les textes ?

Contexte applicatif

- Utilisation des entités nommées
  - **Indexation et recherche d'informations**
  - **Question - réponse**
  - **Annotation en rôles sémantiques**
  - **Résolution d'autres tâches** (transcription, syntaxe, anaphores)
  - …

Contexte applicatif

- Utilisation des entités nommées
  - **Indexation et recherche d'informations**
  - **Question - réponse**
  - **Annotation en rôles sémantiques**
  - **Résolution d'autres tâches** (transcription, syntaxe, anaphores)
  - …

- Collecter des informations sur les entités nommées
  - **Détection** : indiquer où sont les entités nommées
  - **Reconnaissance** : préciser leur type (personne, ville, société, etc.)
  - **Résolution** : déterminer toutes leurs propriétés (référence, valeur)

Que sont les entités nommées ?

- Aux origines : campagnes MUC (DARPA, 1995-2000)

Que sont les entités nommées ?

- Aux origines : campagnes MUC (DARPA, 1995-2000)
- Pas de définition stable
  - « Noms propres et quantités d'intérêt » (MUC)
  - « Désignateurs rigides d'entités » (Kripke)
  - « Entités du monde concret qui ont un nom » (ESTER)
  - « Expression qui réfère à une entité unique » (Ehrmann)
  - …

Que sont les entités nommées ?

- Aux origines : campagnes MUC (DARPA, 1995-2000)
- Pas de définition stable
  - « Noms propres et quantités d'intérêt » (MUC)
  - « Désignateurs rigides d'entités » (Kripke)
  - « Entités du monde concret qui ont un nom » (ESTER)
  - « Expression qui réfère à une entité unique » (Ehrmann)
  - …

$\Rightarrow$ Définitions **applicatives** ou en **extension**

Que sont les entités nommées ?

- Aux origines : campagnes MUC (DARPA, 1995-2000)
- Pas de définition stable
  - « Noms propres et quantités d'intérêt » (MUC)
  - « Désignateurs rigides d'entités » (Kripke)
  - « Entités du monde concret qui ont un nom » (ESTER)
  - « Expression qui réfère à une entité unique » (Ehrmann)
  - …

$\Rightarrow$ Définitions **applicatives** ou en **extension**

- Proposition : objets **à la frontière de la logique**
  « Les entités nommées, lorsqu'elles sont résolues, semblent désigner des objets mentaux de manière stable, à partir desquels il est attendu qu'une représentation logique opère »

Les entités nommées pour ... tous !

Les entités nommées pour ... tous !

Google

quel âge a Maître Gims ?

Connexion

Tous Actualités Vidéos Images Shopping Plus Paramètres Outils

Environ 124 000 résultats (0,52 secondes)

Maître Gims / Âge

30 ans

6 mai 1986

Black M
31 ans

Soprano
37 ans

Kendji Girac
20 ans

Commentaires

Maître Gims — Wikipédia
https://fr.wikipedia.org/wiki/**Maître_Gims**
Maître Gims, de son vrai nom Gandhi Djuna, est un rappeur, chanteur et compositeur congolais né le **6 mai 1986** à Kinshasa. Il est issu d'une famille de ...

Maître Gims

Rappeur

Disponible sur

YouTube
Spotify
Deezer
TuneIn

Maître Gims, de son vrai nom Gandhi Djuna, est un rappeur, chanteur et compositeur congolais né le 6 mai 1986 à Kinshasa. Wikipédia

Naissance : 6 mai 1986 (30 ans), Kinshasa, République démocratique du Congo

Entités nommées : traitements TAL

- Quelques phénomènes linguistiques en jeu

Entités nommées : traitements TAL

- Quelques phénomènes linguistiques en jeu
  - Synonymie
    « **Gandhi Djuna** »
    - $\Rightarrow$ Résolu par l'utilisation de lexiques
    - $\Rightarrow$ Augmente les cas homonymes

Entités nommées : traitements TAL

- Quelques phénomènes linguistiques en jeu
  - Synonymie
    « **Gandhi Djuna** »
    - ⇒ Résolu par l'utilisation de lexiques
    - ⇒ Augmente les cas homonymes
  - Homonymie
    « Il est venu à **Charles de Gaulle**
    - ⇒ Ne peut-être résolue qu'en contexte
    - ⇒ Référence ambigüe

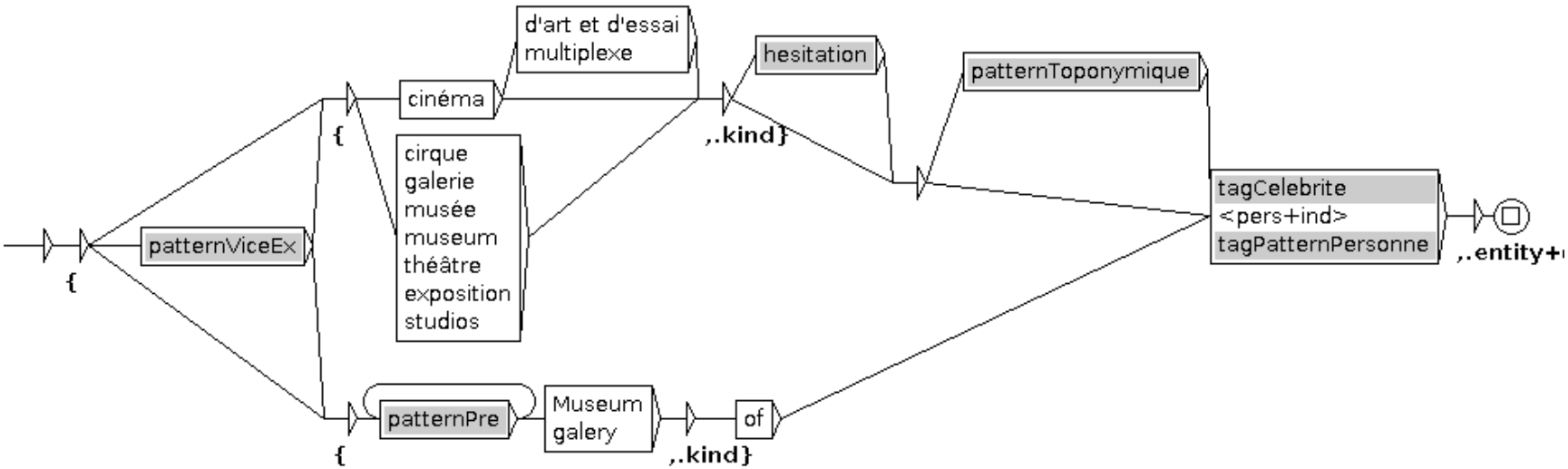
Entités nommées : traitements TAL

- Quelques phénomènes linguistiques en jeu
  - Synonymie
    « **Gandhi Djuna** »
    - ⇒ Résolu par l'utilisation de lexiques
    - ⇒ Augmente les cas homonymes
  - Homonymie
    « Il est venu à **Charles de Gaulle**
    - ⇒ Ne peut-être résolue qu'en contexte
    - ⇒ Référence ambigüe
  - Métonymie « **Paris** a gagné contre **Metz** … »
    - ⇒ Interprétation du référent

Approches orientées connaissances

- Approche très répandue à base d'automates ou de transducteurs

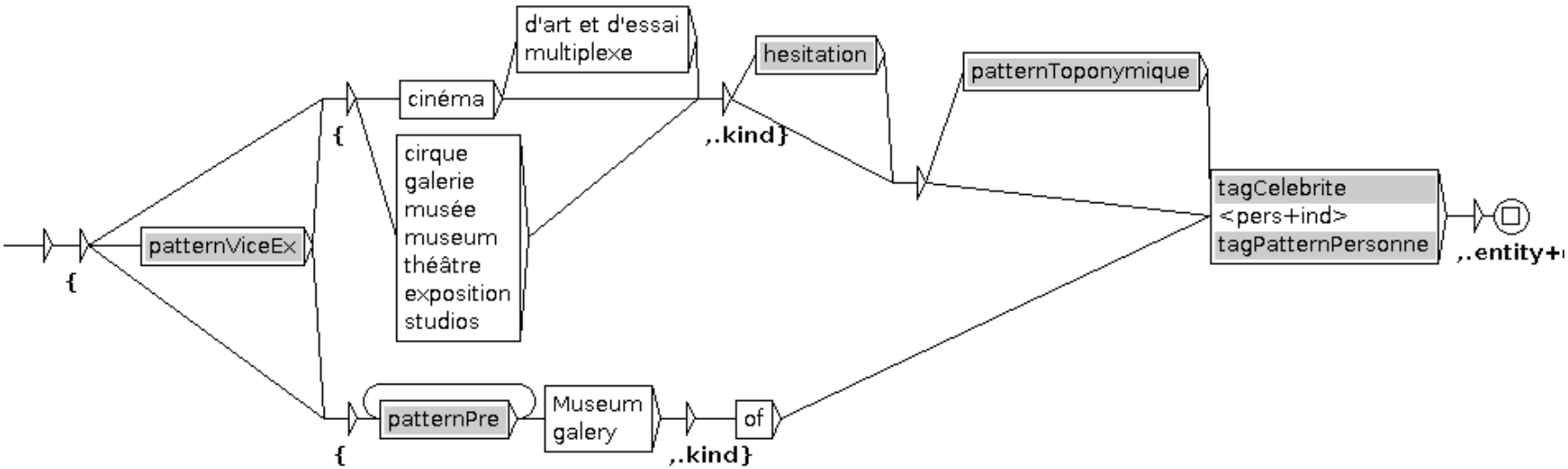
  [McDonald 1996, Friburger 2002, Stern & Sagot 2010, Maurel et.al. 2011]



Approches orientées connaissances

- Approche très répandue à base d'automates ou de transducteurs

  [McDonald 1996, Friburger 2002, Stern & Sagot 2010, Maurel et.al. 2011]



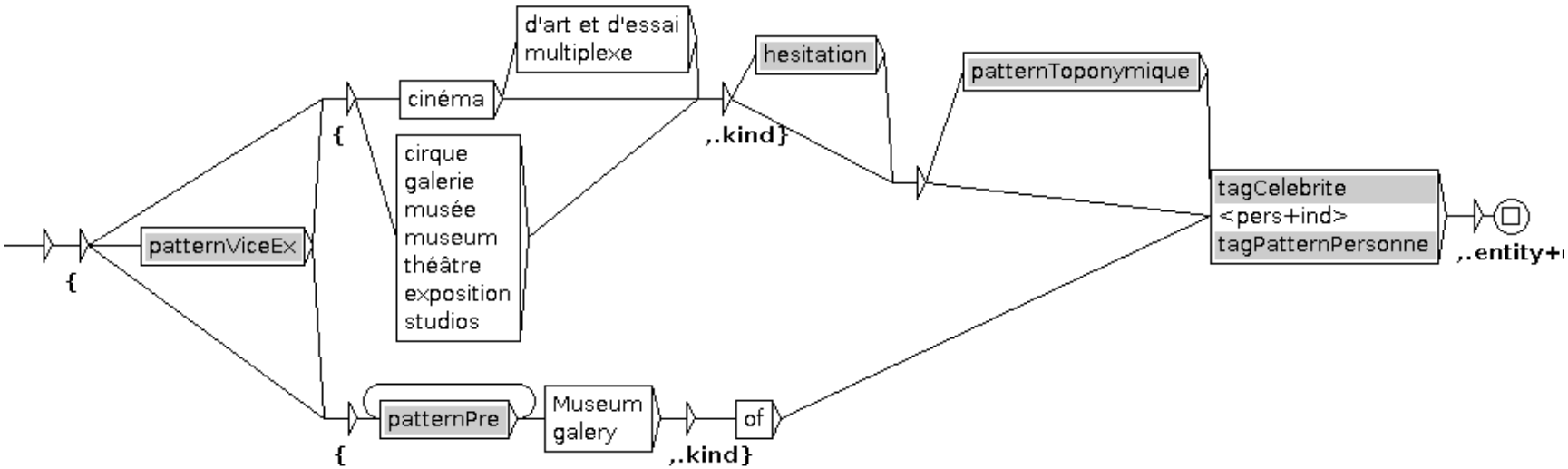
⇒ + Très **précise** (contexte) et **contrôlée**

⇒ + Connaissance capitalisée

Approches orientées connaissances

- Approche très répandue à base d'automates ou de transducteurs

  [McDonald 1996, Friburger 2002, Stern & Sagot 2010, Maurel et.al. 2011]



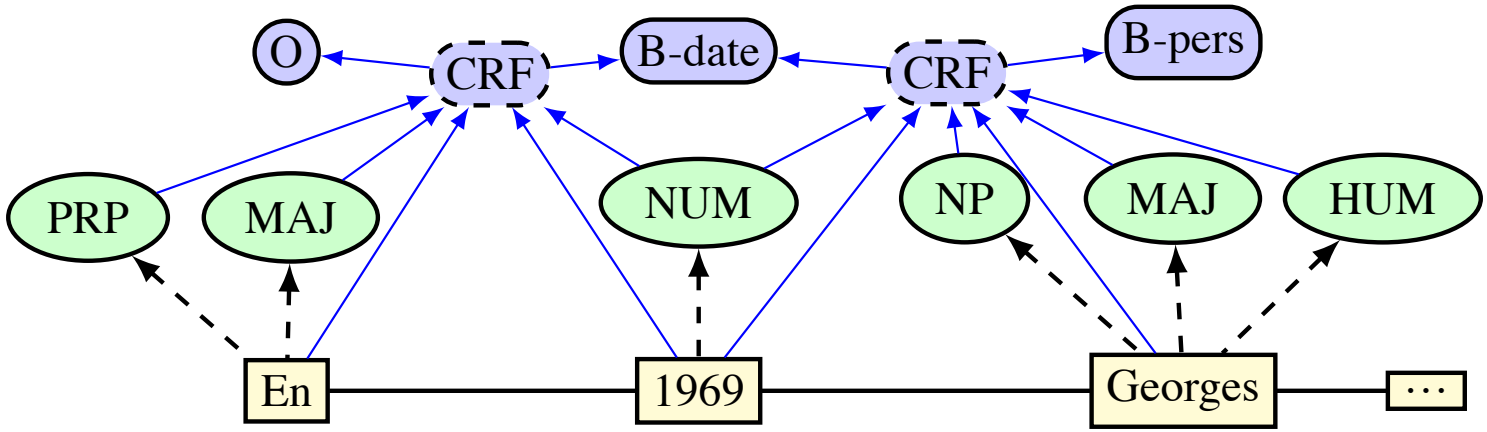
⇒ + Très **précise** (contexte) et **contrôlée**

⇒ + Connaissance capitalisée

⇒ - **Coûteuse** à développer (experts), difficile à adapter

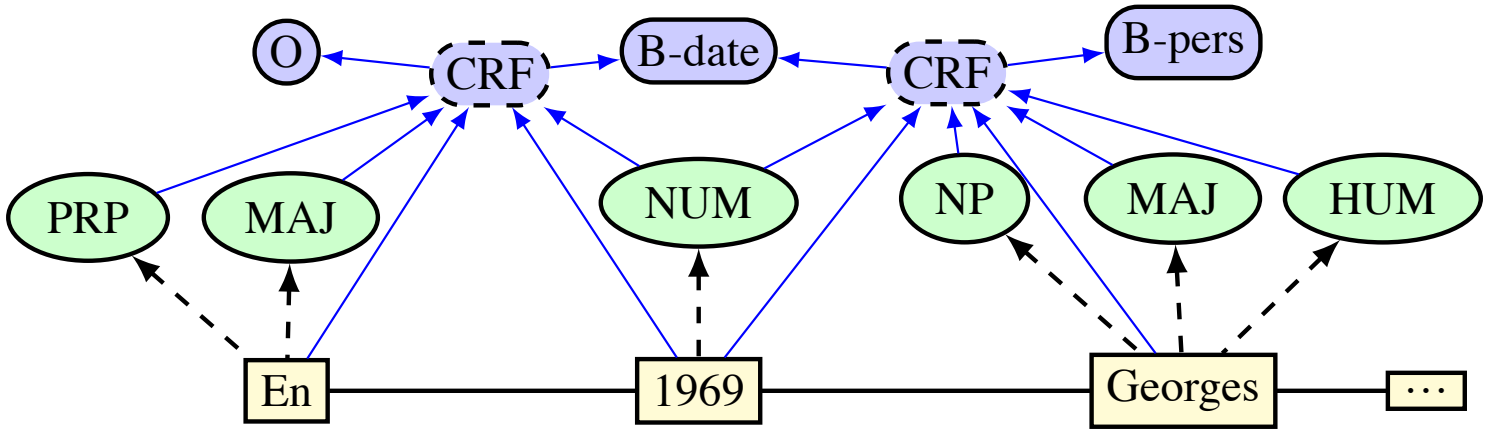
Approches orientées données

- Estimation de la probabilité pour un mot d'être annoté comme une entité nommée (classification) [Borthwick et. al 1998, Mikheev 1999, McCallum 2000, Raymond & Fayolle 2010]
  - Fonctions caractéristiques (régression logistique)
  - Transitions sur la séquence (HMM)



Approches orientées données

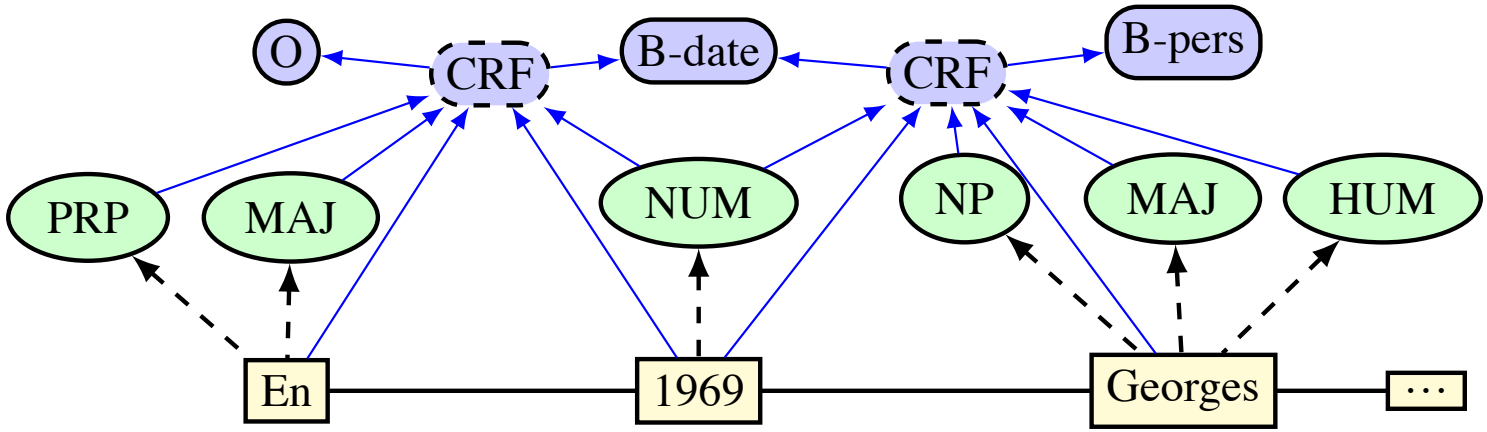
- Estimation de la probabilité pour un mot d'être annoté comme une entité nommée (classification) [Borthwick et. al 1998, Mikheev 1999, McCallum 2000, Raymond & Fayolle 2010]
  - Fonctions caractéristiques (régression logistique)
  - Transitions sur la séquence (HMM)



$\Rightarrow$ + **Automatique** et **robuste**

Approches orientées données

- Estimation de la probabilité pour un mot d'être annoté comme une entité nommée (classification) [Borthwick et. al 1998, Mikheev 1999, McCallum 2000, Raymond & Fayolle 2010]
  - Fonctions caractéristiques (régression logistique)
  - Transitions sur la séquence (HMM)



⇒ + **Automatique** et **robuste**

⇒ - Décisions locales mot-à-mot

⇒ - Plus difficile à interpréter, **adapter** et capitaliser

mXS

- Logiciel REN
  - Doctorat (2012) « Extraction automatique de motifs »
  - Annotation par **marqueurs** (segmentation plus qu'étiquetage)
  - **Enrichissement ambigu** : POS, ressources (lexiques, automates)
  - **Désambiguisation statistique** : régression logistique
  - Dispo en beta : résolution des entités pour une base ou Wikipedia

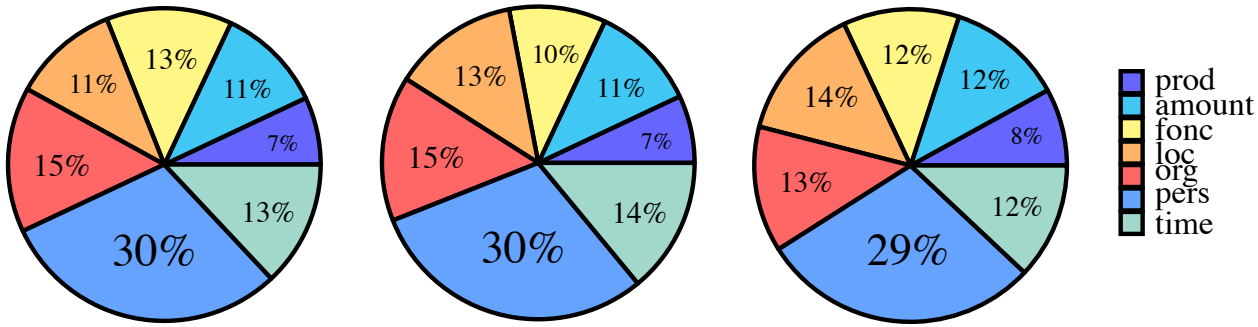
mXS

- Logiciel REN
  - Doctorat (2012) « Extraction automatique de motifs »
  - Annotation par **marqueurs** (segmentation plus qu'étiquetage)
  - **Enrichissement ambigu** : POS, ressources (lexiques, automates)
  - **Désambiguisation statistique** : régression logistique
  - Dispo en beta : résolution des entités pour une base ou Wikipedia
- En pratique
  - Deux modèles
    - Etape : entités et composants étapes (imbrications)
    - EtapePLOP : « grossier » (personnes, lieux, organisations, produits)
  - Disponible en ligne : `https://github.com/eldams/mXS`
  - Peu d'évolution depuis 2012 (lexiques datés à MAJ un jour)
  - Toutes contributions bienvenues …

ETAPE : corpus

- 74 enregistrements annotés (BFMTV, France Inter, LCP, TV8)

Corpus	Tokens	Enoncés	EN
Etape-Train	355 975	14 989	46 259
Etape-Dev	115 530	5 724	14 112
Etape-Test	123 221	6 770	13 055
Total	594 726	27 483	73 426



(a) Etape-Train (b) Etape-Dev (c) Etape-Test

$\Rightarrow$ 3 entités par énoncé / 1 entité tous les 7 mots

ETAPE : résultats

⇒ ETAPE, systèmes orientés connaissances / orientés données

Part.	**Manuel**	**Rover**	**WER 23**	**WER 24**	**WER 25**	**WER 30**	**WER 35**
S1	85,6	98,1	100,7	94,2	98,9	98,4	100,9
S2	156,6	147,4	178,8	160,4	168,0	163,9	168,2
S3	36,6	**57,2**	**59,3**	64,7	**62,0**	**61,7**	**71,8**
S4	50,5	88,0	98,8	76,8	92,8	94,9	99,6
S5	44,8	69,7	73,8	72,1	73,7	74,8	86,0
S6	na	79,2	79,5	66,8	80,8	80,0	87,0
S7	na	67,8	68,4	67,6	70,9	69,9	85,2
S8	37,5	na	na	na	na	na	na
S9	62,5	75,8	79,2	76,9	79,8	80,5	90,5
S10	39,3	65,0	69,9	66,3	70,5	69,9	87,0
CasEN	**35,3**	na	na	68,4	na	na	na
mXS	38,4	63,7	67,5	**64,1**	69,1	68,6	80,4
Position	4	2	2	**1**	2	2	2

Résultats CAP 2017

- `http://cap2017.imag.fr/competition.html`
- Tweets, personnes, lieux, produits, équipes, transports, etc.

Rang	Système	Mesure-F	Précision	Rappel
1	Synapse Développement	**58.89**	73.65	**49.06**
2	HIT - Agadir	52.19	58.95	46.83
3	**TanDam**	51.99	60.67	45.48
4	**NER Quebec**	51.26	67.65	41.26
5	Swiss Chocolate	50.05	56.42	44.97
6	AMU-LIF	46.21	53.59	40.63
7	Lattice	45.46	**78.76**	31.95
na	mXS	37.97	59.55	27.87
8	Geolsemantics	21.28	19.66	23.18

Résoudre /lier les entités (nommées)

- Défi en cours de résolution pour les entités (nommées)
  - Reconnaissance insuffisante : *M. Hollande*
  - Liaison avec le **web sémantique** (LOD)

  $\Rightarrow$ Reconnaissance de **mentions** d'entités

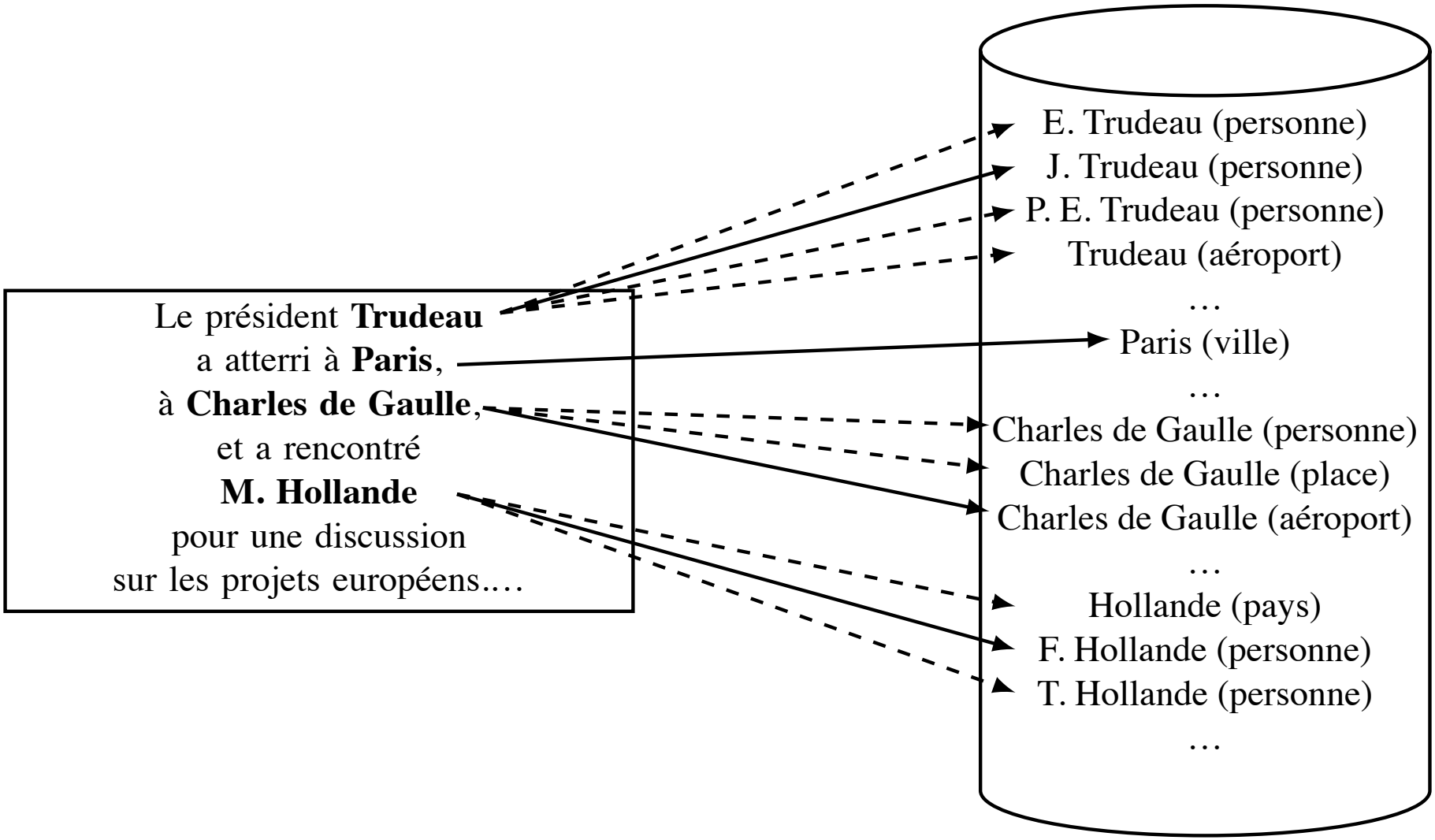
Résoudre /lier les entités (nommées)

- Défi en cours de résolution pour les entités (nommées)
  - Reconnaissance insuffisante : *M. Hollande*
  - Liaison avec le **web sémantique** (LOD)

  $\Rightarrow$ Reconnaissance de **mentions** d'entités
- Exploitation d'un **modèle** / référentiel
  - Liste aussi large que possible d'entités d'intérêt
  - Nécessairement **incomplet** (`NIL`)
  - Liaison : coréférences, anaphores, clustering …

$\Rightarrow$ Exploitation intensive de **Wikipedia**

Lier les textes aux bases de connaissances



Principes généraux de résolution

- Prétraitement par **reconnaissance** rudimentaire
  - Insensibilité à la casse (majuscules / minuscules)
  - Variantes liées aux jeux de caractères (diacritiques, liaisons, etc.)
  - Suppression des mots-outils / de parenthèses
  - Génération automatique d'acronymes à partir des formes

$\Rightarrow$ De nombreuses liaisons potentielles

Principes généraux de résolution

- Prétraitement par **reconnaissance** rudimentaire
  - Insensibilité à la casse (majuscules / minuscules)
  - Variantes liées aux jeux de caractères (diacritiques, liaisons, etc.)
  - Suppression des mots-outils / de parenthèses
  - Génération automatique d'acronymes à partir des formes

$\Rightarrow$ De nombreuses liaisons potentielles

- La **désambiguisation** contextuelle
  - **Texte** du document : mots, domaine, autre entités (anaphores)
  - **Popularité** du référent
  - Attributs du référent : type, nom / titre (redirections), description

$\Rightarrow$ Apprentissage (SVM) et classification des référents par mention

Quelques résultats de résolution

- Performances (anglais, textes généraux) : 80% à 90%

Quelques résultats de résolution

- Performances (anglais, textes généraux) : 80% à 90%
- Nombreux logiciels en ligne, dont
  - BabelFy (BabelNet)
  - Wikipedia Spotlihght
  - Alchemy
  - The Wiki Machine
  - NERD
  - …

$\Rightarrow$ Résolveurs par domaine (juridique, historique, biomédical, etc.)

Plan

http://damien.nouvels.net/bazar/desamb.pdf

La désambiguisation totale, un Graal

- **Désambiguisation** comme **compréhension**
  - Qui a dit que l'humain « comprend » ?!

  $\Rightarrow$ Comprendre aussi « moins mal possible » que l'humain …

  $\Rightarrow$ Accords inter-annotateurs : entités, opinion, traduction

  $\Rightarrow$ Ne désambiguiser que ce qui peut **raisonnablement** l'être

  $\Rightarrow$ Mais garder aussi de l'ambiguïté …

La désambiguisation totale, un Graal

- **Désambiguisation** comme **compréhension**
  - Qui a dit que l'humain « comprend » ?!

  ⇒ Comprendre aussi « moins mal possible » que l'humain …

  ⇒ Accords inter-annotateurs : entités, opinion, traduction

  ⇒ Ne désambiguiser que ce qui peut **raisonnablement** l'être

  ⇒ Mais garder aussi de l'ambiguïté …
- Méthodes de **reconnaissance / résolution**
  - Opérationnelles au niveau des ambiguïtés locales
  - Peu de prise en compte inter-phrastique ou des méta-informations

  ⇒ Réseau de neurones ou pas …

  ⇒ Quelles méthodes plus globales ?

La désambiguisation totale, un Graal

- **Désambiguisation** comme **compréhension**
  - Qui a dit que l'humain « comprend » ?!

  ⇒ Comprendre aussi « moins mal possible » que l'humain …

  ⇒ Accords inter-annotateurs : entités, opinion, traduction

  ⇒ Ne désambiguiser que ce qui peut **raisonnablement** l'être

  ⇒ Mais garder aussi de l'ambiguïté …
- Méthodes de **reconnaissance / résolution**
  - Opérationnelles au niveau des ambiguïtés locales
  - Peu de prise en compte inter-phrastique ou des méta-informations

  ⇒ Réseau de neurones ou pas …

  ⇒ Quelles méthodes plus globales ?
- Adaptation indispensable (et coûteuse) aux **langues**

La désambiguisation totale, un Graal

- **Désambiguisation** comme **compréhension**
  - Qui a dit que l'humain « comprend » ?!

  ⇒ Comprendre aussi « moins mal possible » que l'humain …

  ⇒ Accords inter-annotateurs : entités, opinion, traduction

  ⇒ Ne désambiguiser que ce qui peut **raisonnablement** l'être

  ⇒ Mais garder aussi de l'ambiguïté …
- Méthodes de **reconnaissance / résolution**
  - Opérationnelles au niveau des ambiguïtés locales
  - Peu de prise en compte inter-phrastique ou des méta-informations

  ⇒ Réseau de neurones ou pas …

  ⇒ Quelles méthodes plus globales ?
- Adaptation indispensable (et coûteuse) aux **langues**
- **Exigences** : objectivité des évaluations, méthodes reproductibles

HackaTAL 2017

- **Hackathon** dans le domaine du **TAL**
- Évènement / atelier TALN
- **2016**
  - **Google** (Paris)
  - 50 personnes
  - Tâches : chatbots / détection d'évènements (foot)
- **2017**
  - **Lab'O** (Orléans)
  - **Tâches**
    - *Résumé auto. de commentaires de produits en ligne (hôtellerie)*
    - *Identification des tendances stratégiques liées aux brevets*

$\Rightarrow$ `http://hackatal.github.io/2017/`

$\Rightarrow$ Gratuit, tout le monde est bienvenu - inscrivez-vous !

Merci !

« Bella »

15 pers : 13 « Bella », 2 « Juste »

⇒ Focalisé sur une personne …

« Sapés comme jamais »

5 loc : 1 « Abidjan », 1 « Bamako », 1 « Libreville », 2 « Nuits Paris »
8 org : 6 « Chanel », 1 « Gestapo », 1 « Hermès Louis Vuitton »
9 pers : 3 « Chanel », 1 « Charlie », 1 « Congolais », 1 « Dany Synthé », 1 « Darcy », 1 « Djuna Djanana », 1 « Gustavo »

⇒ Très africain et mode

« Est-ce que tu m'aimes ? »

Rien …

⇒ Trop …« lyrique » ?!

`http://damien.nouvels.net/bazar/desamb.pdf`