LABORATOIRE D'INFORMATIQUE GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

4/7/2014
Informatique et SHS
Rencontre INRA-INRIA
Paris

Utilisation informatique de résultats

Un défi pour la linguistique

Éric Laporte

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

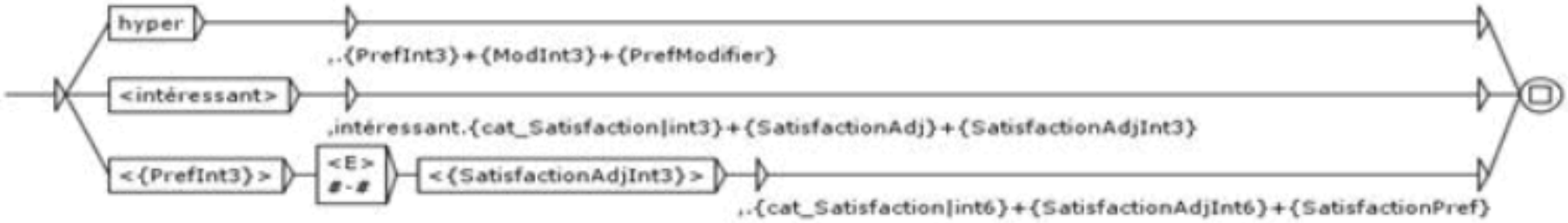
Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Sommaire

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Thème de réflexion



Source : Martineau *et al.*, 2014

Résultats de la linguistique exploitables pour le traitement des langues (ressources) :

- grammaires
- dictionnaires
- corpus annotés
- ontologies

En produire est un défi pour la linguistique

Comment peut-elle le relever ?

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Sommaire

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Défis

Concurrence avec l'apprentissage automatique

Contrôle de qualité

Formalisation

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Concurrence avec l'apprentissage automatique

Dans un moule à tarte rond antiadhésif, faire fondre directement sur le feu le beurre. Ajouter dans le plat le sucre fin, et baisser le feu pour faire un caramel.	In a round pie pan nonstick, melt over direct heat butter. Add the flat end of the sugar and reduce heat to a caramel.

Source : Google Translate

Traduction automatique statistique	corpus bilingue
Analyse syntaxique probabiliste	corpus annoté
Acquisition de dictionnaires syntaxiques	corpus annoté
Acquisition d'ontologies	corpus

L'apprentissage automatique vise à se passer de dictionnaires et de grammaires

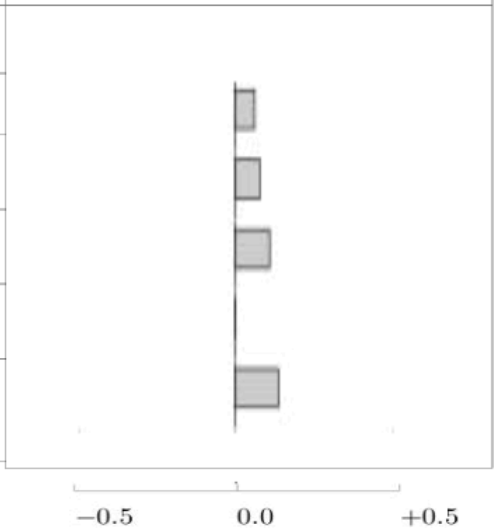
Activité de même nature : généralisation à partir d'exemples
Puissance de calcul différente
Facultés différentes

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Contrôle de qualité

Méthode	Amb.	Tagset	%noms annotés	F-mesure/ Etiquetage	Gains absolus (F-mesure)
Baseline	-	28	-	85.05/97.43	
AnnotTable	1	98	8,6%	85.10/97.42	
AnnotNoms	1	33	11,2%	85.13/97.48	
AnnotNoms	2	38	16,5%	85.16/97.47	
AnnotNoms	3	39	16,9%	85.05/97.41	
AnnotIN	-	30	16,9%	**85.20/97.54**	



Source : Sigogne *et al.*, 2014

En traitement des langues, on mesure les performances

Elles peuvent dépendre de la fiabilité et de l'exhaustivité des ressources

Les informaticiens reprochent aux linguistes de ne pas décrire l'usage réel

Distance culturelle : en linguistique, des commentaires intéressants sont traditionnellement un résultat en soi

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Formalisation

réduire/N0 : chirurgien/N1 : fracture/N2 :/S: rebouter/A:
réduire /N0 : hum/N1 : minerai/N2 :/S: éliminer l'oxygène de/A:
réduire /N0 : hum/N1 : (sauce, jus)/N2 :/S: épaissir/A: allonger
réduire /N0 : hum/N1 : fils/N2 :/S : rapprocher/A: écarter
réduire /N0 : hum, pays/N1 : hum, pays/N2:/S : vaincre/A:libérer

réduire /N0 : hum /N1 : hum/N2 : en <esclavage>/S: rabaisser/A: sortir
réduire /N0 : hum, évé/N1 : hum/N2 : à <état>/S : contraindre/A:
réduire /N0 : hum, évé/N1 : hum/N2 : à <action>/S : contraindre/A: libérer
réduire /N0 : hum /N1 :<tout> /N2 : à <Npt >/S : diviser/A: recomposer
réduire /N0 : hum /N1 :inc/N2 : en <miettes, pièces>/S : casser/A: recoller

réduire /N0 : photographe/N1 : photo/N2 : de %/S: diminuer/A: agrandir
réduire /N0 : hum/N1 :<valeur>/N2 : de %/card /S : diminuer/A: augmenter
réduire /N0 : hum/N1 :< un texte>N2 : de %/S: raccourcir/A:

Source : Gross, 2008

Historiquement, la linguistique fait preuve d'une résistance à la formalisation

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Sommaire

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Solutions

Annotation de corpus et révision

Modélisation

Critères formels

Couverture lexicale et grammaticale

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Annotation de corpus et révision

Prefrontal cortex in the rat: projections to subcortical autonomic, motor, and limbic centers.
This paper describes the quantitative areal and laminar distribution of identified neuron populations projecting from areas of prefrontal cortex (PFC) to subcortical autonomic, motor, and limbic sites in the rat. Injections of the retrograde pathway tracer wheat germ agglutinin conjugated with horseradish peroxidase (WGA-HRP) were made into dorsal/ventral striatum (DS/VS), basolateral amygdala (BLA), mediodorsal thalamus (MD), lateral hypothalamus (LH),

Source : French *et al.*, 2009

Avantages

Facile à utiliser par l'apprentissage automatique
Analyse d'exemples réels
Confrontation avec la réalité

Inconvénients

Travail répétitif
Ne met pas à contribution la capacité humaine à généraliser
Le linguiste est sous-utilisé

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Modélisation

- France fell into recession. Pulled out by Germany.
- US Economy on the verge of falling back into recession after moving forward on an anemic recovery.

Source : Narayanan, 2012

Métaphores de l'espace vers des concepts abstraits

Modèle psycholinguistique

Reconstitution du fonctionnement mental du locuteur

Modèle purement linguistique

Entrées lexicales distinctes

N_0 fall Loc N_1	*A man fell onto the tracks*
N_0 Vsup recession	*France (had a + was in + came into + fell into) recession*
N_0 Vsup verge	*The lane has a wide verge*
N_0 Vsup on the verge of N_1	*I'm on the verge of crying*

Origine : linguistique structurale

Objets mieux définis que les processus mentaux
Met à contribution la capacité humaine à comparer les sens
Objets porteurs de sens plus précis que les mots (*fall, verge*)

Défi pour le traitement automatique : objets complexes

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Critères formels

La tension du malade est élevée
Le prix de ce sac est modique, dérisoire
La séance est courte
Le salaire de Luc est ridicule, confortable
La dénivellation est forte

Source : Giry-Schneider, 2011

Dérisoire décrit la quantité avec les noms de grandeurs

Toute cette histoire est dérisoire
Le prix de ce sac est dérisoire

Le prix de ce sac est de combien ? — Il est de 30 euros
**Toute cette histoire est de combien ? — Elle est de Dnum N*

Origine : linguistique distributionnelle

Reproductibilité des observations
Fiabilité des ressources
Met à contribution la capacité humaine à comparer les sens

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Couverture lexicale et grammaticale

Adj	Prép	Exemple	Nq de N être Adj = N être Adj	Dét Adj N0	très Adj	Nq = quantité	Nq = niveau	Nq = montant	Nq = durée	N0 être plus Adj que N0	N0 être plus Adj de Dnum unités que N0	Npréd de N0 être Adj	Adj-n
abondant		La récolte de blé est abondante	+	+	+	+	–	–	–	+	+	–	abondance
abordable		Le prix du blé est abordable	+	–	+	–	–	+	–	+	–	–	–
abyssal		L'écart entre ces deux sommes est abyssal	–	–	–	–	–	+	–	–	–	+	–
accablant		Ce niveau de chaleur est accablant	+	–	–	–	+	–	–	+	–	–	–
acceptable		Le prix de ce livre est acceptable	–	–	+	+	+	+	–	+	–	+	–
affligeant		Cette quantité de blé est affligeante	–	+	–	+	+	+	–	–	–	+	–
affolant		Le prix du tabac est affolant	–	–	–	+	+	+	–	–	–	+	–
ahurissant		Cette quantité de blé est ahurissante	–	+	–	+	+	+	–	–	–	+	–
ample		L'oscillation de ce pendule est ample	–	+	+	–	–	–	–	+	+	–	amplitude

Origine : lexique-grammaire (Gross, 1981)

Balayage descriptif

Confrontation avec la réalité : expressions polylexicales, constructions à verbe support

Défi pour le traitement automatique : sélectionner les entrées pertinentes pour une application

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Que retenir ?

Modélisation, reproductibilité, couverture

Des notions liées à la scientificité

La linguistique a des moyens de relever les défis du traitement automatique des langues

Des moyens plutôt vintage

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Merci

CONTACT

ÉRIC LAPORTE

00 +33 (0)1 60 95 75 52

ERIC.LAPORTE@UNIV-PARIS-EST.FR