LABORATOIRE D'INFORMATIQUE GASPARD-MONGE

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

2014/09/04
Computational Linguistics in Bulgaria
Sofia

Interaction between Linguists and Machine Learning

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

Topic

лодка,N602+f
лондонски,A2
мъж,N4+m
параход,N8+m
Париж,N7+m+Nprop
плавателен,A5
съд,N1+m
фракция,N603+f
франция,N601+f+NProp
французин,N9+m+NProp
французки,A2
червен,A3
член,N5+m
човек,N6+m

Source: Svetla Koeva, Cvetana Krstev

Language resources for language processing:

- grammars
- dictionaries
- annotated corpora
- ontologies

Producing usable resources is a challenge to us descriptive linguists

What are our strong points?

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Outline

Three challenges to linguists

Which solutions

Conclusions

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Three challenges

Competing with machine learning

Facing quality control

Formalizing

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Competing with machine learning

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

Statistic-based translation	bilingual corpus
Probabilistic syntactic parsing	annotated corpus
Syntactic dictionary acquisition	annotated corpus
Ontology acquisition	corpus

Machine learning was designed to dispense with dictionaries and grammars

Same type of activity

Generalization from examples

If I describe the behaviour of *plat,* I base myself on examples

Which performs better?

Computational power

Linguists have, for example, an ability to compare meanings: *plat* "flat", "dish"

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Facing quality control

Language processing module	Precision %	Recall %	Evaluation data
sentence splitter	92.00	99.00	190 sentences
paragraph splitter	94.00	98.00	268 paragraphs
clause chunker	93.50	93.10	232 clauses
POS tagger	95.00	95.00	303 POS tags
NP extractor	63.50	77.00	352 NPs

Source: Tanev & Mitkov, 2002

In language processing, we test applications for performance

Testing language resources for quality

Reliability

Coverage (or exclusivity to the domain)

Performance of applications

Quality is not easy to achieve

Computer scientists complain that linguists are purists, do not describe real-world usage

Cultural distance

Linguistics lacks a tradition of quality control

Interesting comments are traditionally a result *per se*

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Formalizing

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

Identified fields; no texts (definitions or examples)
Historically, linguistics resists to formalization
Argument classes are represented by lemmas: *photo*, sequences: *un texte* "a text",
sequences with inflected words: *en miettes* "into pieces", codes: hum

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Outline

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Linguists' assets

What abilities allow us linguists to take on these challenges?
- Do corpus annotation and revision
- Create and use models
- Select relevant goals
- Apply formal criteria
- Extend lexical and grammatical coverage of resources

Which trends prepares us best?

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Corpus annotation and revision

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

The dominant model of interaction between linguists and machine learning

Easy to use for machine learning

Analysis of real examples

Confrontation with the real world

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Corpus annotation and revision

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

The 'easy option'

Only apparently satisfactory

Repetitive work
- Does not make full use of human ability to generalize
- Linguist is under-employed
- Who likes annotating a corpus?

Some information is usually missing
- Identifiers of lexical entries in case of lexical ambiguity
- Identifiers of syntactic constructions

These issues are specific to annotated corpora

We have other weapons in our arsenal

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Identifiers of lexical entries in case of lexical ambiguity

There is water under the sea floor	noun	
Our neighbour will water the garden	verb	2 entries
You packed your own luggage	no *with*-arg	
The house was packed with art works	*with*-arg	2 entries
Ann announced her pregnancy	no *to*-arg	
Ann announced her pregnancy to the public	*to*-arg	same entry

No feature or combination of features is equivalent to the information of whether 2 occurrences belong to a single lexical entry

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Creating and using models

- 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

Spatial metaphors of abstract concepts

We represent phenomena within models

Psycholinguistic model
Mental processes of language users

Purely linguistic model with lexical entries
Conventional metaphors: distinct lexical entries

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*

Linguistic forms are easier to observe than mental processes
Origin: structural linguistics

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Creating and using models

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*

Models with lexical entries

As compared to corpus annotation

- Make full use of human ability to compare meanings
- Lexical entries represent more accurate meanings than words (*fall, verge*)
- Challenge to language processing: complex objects

But lexical entries make sense as elements of a formal model

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Selecting relevant goals

Example: inventorying arguments of predicates

Goal 1: assign each argument a semantic role

John opened the door — *The door opened*
Agent — Patient — Patient

Students like social media
Experiencer — Causer? Theme? Stimulus?

Goal 2: number each argument (Gross, 1975, 1994)

John opened the door — *The door opened*
N_0 — N_1 — N_1

Students like social media
N_0 — N_1

Neither goal has been fulfilled yet, even for the most studied languages

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Selecting relevant goals

Goal 1: qualify each argument with a semantic role

Students like *social media*

Experiencer Causer? Theme? Stimulus?

Goal 2: number each argument

Students like *social media*

N_0 N_1

Comparison as regards use in applications

Goal 2 is sufficient to identify the arguments of a predicate

This is what is required for translation, information extraction...

Other benefits of goal 1 are hypothetical

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Selecting relevant goals

Goal 1: qualify each argument with a semantic role

Students like *social media*

Experiencer Causer? Theme? Stimulus?

Goal 2: number each argument

Students like *social media*

N_0 N_1

Comparison as regards accuracy

Goal 1 has no decisive criteria for distinguishing semantic roles

Majority vote among annotators, crowdsourcing

Goal 2 involves inventorying and arbitrary numbering: practicable

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Selecting relevant goals

Photo: David Whitehorse

Crowdsourcing for semantic role labelling

Influence of syntax is a major pitfall of semantic role labelling

They talked me into this project
Agent Patient Goal

into, locative preposition, therefore goal, a spatial role

Snow covers the car
Agent Patient

'The subject is the doer of the action' (primary school)

Volunteers are most likely to fall into these pitfalls

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Selecting relevant goals

Goal 2 is more useful and more accurately defined

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Applying formal criteria

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

Adjectives describing quantity in French

Dérisoire "derisive" describes quantity with quantity nouns
Toute cette histoire est dérisoire "All this stuff is derisive"
Le prix de ce sac est dérisoire "The price of this bag is miniature"

What is a quantity noun?
Le prix de ce sac est de combien ? — Il est de 30 euros
"What amount is the price of this bag? — It is 30 euros"
**Toute cette histoire est de combien ? — Elle est de* Dnum N
*"What amount is all this stuff? — It is *Dnum N*"

With a formal criterion, recognition of a quantity noun depends less on the observer
Origin: distributional linguistics

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Applying formal criteria

What is a quantity noun?

Le prix de ce sac est de combien ? — Il est de 30 euros

"What amount is the price of this bag? — It is 30 euros"

**Toute cette histoire est de combien ? — Elle est de* Dnum N

*"What amount is all this stuff? — It is *Dnum N*"

Methods with formal criteria

As compared to semantic intuition

- Make full use of human ability to compare meanings
- Reproducibility of observation
- Resource reliability

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Extending coverage

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

Descriptive scan
Origin: lexicon-grammar (Gross, 1975, 1994)
As compared to corpus annotation

Confrontation with the real world
Dictionaries of multiword expressions
Grammars of support-verb constructions
Rare uses of words and rare words

Challenge to language processing
Select entries relevant to an application
But it makes sense to be able to do so

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Outline

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Conclusion

4 notions related to scientificity
Models
Accuracy of goals
Reproducibility of observation
Coverage

Linguistics has methodological weapons
to take on the challenges of language processing

Deeply rooted in the history of linguistics
Structural linguistics
Distributional linguistics
Lexicon-grammar
The legacy of these 3 trends has potential for future

What about current fashionable trends of linguistics?

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

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

Thanks

CONTACT

ÉRIC LAPORTE

00 +33 (0)1 60 95 75 52

ERIC.LAPORTE@UNIV-PARIS-EST.FR