

# Migrating LINA Laboratory to Apache UIMA

Stegos Afantenos et Matthieu Vernier

Équipe TALN - Laboratoire Informatique Nantes Atlantique

Vendredi 10 Juillet 2009

- 1 Introduction
- 2 UIMA : a solution for academic research
- 3 Using UIMA in projects : Blogoscopie and Piithie project
- 4 Conclusion : toward a french community around UIMA

# Overview

- 1 Introduction
- 2 UIMA : a solution for academic research
- 3 Using UIMA in projects : Blogoscopie and Piithie project
- 4 Conclusion : toward a french community around UIMA

# Some words on LINA-TALN

- 1 LINA-TALN is a computational linguistics laboratory conducting research in the following domains :
  - Terminology extraction
  - Grammatical acquisition
  - Semantic annotation in texts
  - Discourse analysis
- 2 Natural Language Processing (NLP) Applications : Machine translation, Information extraction, Opinion mining, Plagiarism detection, etc
- 3 LINA-TALN is participating in various projects financed either from regional or national sources (PIITHIE, BLOGOSCOPIE, MILES, etc).

# General problem

Computational linguistic researchers have developed many tools and components ...

- ① **Reusability** – How can we share components?
  - With encapsulated functionality
  - Written in different programming languages
- ② **Interoperability** – How can we integrate (configure) different components in complex NLP pipelines?
  - Without tremendous programming overhead
- ③ **Testability, Maintainability** – How can we test complex NLP pipelines?
  - With easy modifiability (switching parameters)
- ④ **Portability** – How can we adapt complex NLP pipelines to new domains, languages, apps?

# How do researchers write code ?

- Very often they are interested in a "proof of concept"

# How do researchers write code ?

- Very often they are interested in a "proof of concept"
- By consequence, they seldom are interested in the interoperability of their modules with modules from other researchers in the same laboratory.

# How do researchers write code ?

- Very often they are interested in a "proof of concept"
- By consequence, they seldom are interested in the interoperability of their modules with modules from other researchers in the same laboratory.
- Whence the need for a **common platform**.

# A simple example of interoperability problems

## Alice's adventures in wonderland

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do.

### One way of tokenizing

Alice/was/beginning/to/get/very/  
tired/of/sitting/by/her/sister/  
on/the/bank/,/and/of/having/  
nothing/to/do/.

### A second way of tokenizing

Alice	by	having
was	her	nothing
beginning	sister	to
to	on	do
get	the	.
very	bank	
tired	,	
of	and	
sitting	of	

# Overview

- 1 Introduction
- 2 UIMA : a solution for academic research**
- 3 Using UIMA in projects : Blogoscopie and Piithie project
- 4 Conclusion : toward a french community around UIMA

# How UIMA can help ?

- All researchers need to do is clearly define their type system

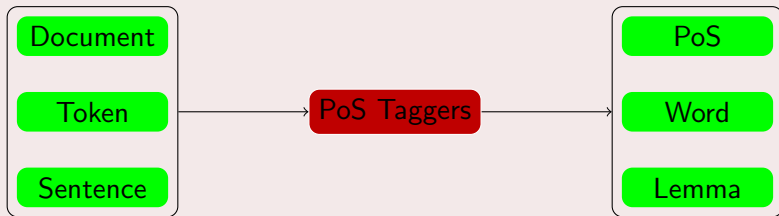
# How UIMA can help ?

- All researchers need to do is clearly define their type system
- Then, each one can go on and write its own code (Analysis Engine, according to the jargon)

# How UIMA can help ?

- All researchers need to do is clearly define their type system
- Then, each one can go on and write its own code (Analysis Engine, according to the jargon)
- UIMA takes care of the rest, allowing for easy chaining of modules, which can be **local** or **distributed** in various machines inside the same laboratory.

## Some annotators (Treetagger, Brill, Flemm)



- The annotation type *Word* is a *meta*-annotation for 25 annotation types, including nouns, pronouns, verbs etc.
- In LINA-TALN we have been working for more than a year now on the definition of a common type system to be used by our team members and maybe by the community at large.

# What about inter-laboratory collaborations ?

- Laboratories very often collaborate with other national or international laboratories

# What about inter-laboratory collaborations ?

- Laboratories very often collaborate with other national or international laboratories
- Nonetheless, they either cannot (due to copyright reasons) or do not want to publicly distribute their code or even executables.

# What about inter-laboratory collaborations ?

- Laboratories very often collaborate with other national or international laboratories
- Nonetheless, they either cannot (due to copyright reasons) or do not want to publicly distribute their code or even executables.
- Solution : implementation of a **web service**

# What about inter-laboratory collaborations ?

- Laboratories very often collaborate with other national or international laboratories
- Nonetheless, they either cannot (due to copyright reasons) or do not want to publicly distribute their code or even executables.
- Solution : implementation of a **web service**
- UIMA, with the **Simple Rest Server** provides an easy to use framework for the exchange on information based on the REST protocol

# What about inter-laboratory collaborations ?

- Laboratories very often collaborate with other national or international laboratories
- Nonetheless, they either cannot (due to copyright reasons) or do not want to publicly distribute their code or even executables.
- Solution : implementation of a **web service**
- UIMA, with the **Simple Rest Server** provides an easy to use framework for the exchange on information based on the REST protocol
- In case the other laboratories use UIMA the SOAP or jVinci protocols are also available.

# UIMA at LINA

- 1 Definition of a basic Type System for french preprocessing NLP task (tokenizer, grammatical tagging)
- 2 Development of UIMA wrappers for existing tools : *TreeTagger*, *Brill*, *Flemm*
- 3 Rewriting of *Nemesis* in UIMA : named entity annotation tool
- 4 Using UIMA in our Masters courses in order to teach “good practices” of NLP development.
- 5 Several academic and industrial projects :
  - ANR CMantic : semantic search engine development
  - ANR Piithie : semantic and discursive analysis by REST web service for plagiarism detection
  - ANR Blogoscopie : opinion mining in blogs

# Overview

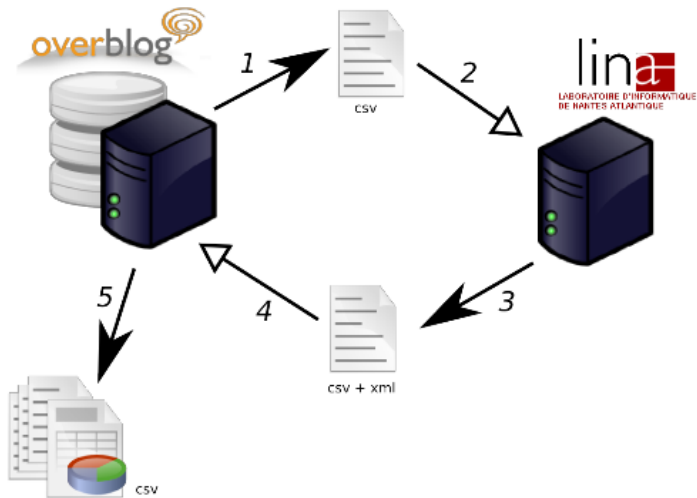
- 1 Introduction
- 2 UIMA : a solution for academic research
- 3 Using UIMA in projects : Blogoscopie and Piithie project
- 4 Conclusion : toward a french community around UIMA

# UIMA in Blogoscopie project

- Objective : *What do peoples think about a subject at a particular date ?*
  - Automatic annotation and categorization of local opinions expressed in blogs
  - Automatic annotation of evaluated subjects
- Industrial partner : OverBlog (1st french blog platform)
- Tools needed for this NLP :
  - Morphological/Syntax annotator (TreeTagger),
  - Dictionary annotator (Opinion words like *interesting, to hate, etc*)
  - Rules annotator (*NEG + ADV + Opinion Word*, e.g. “not very useful”),
  - Machine learning algorithms,
  - etc

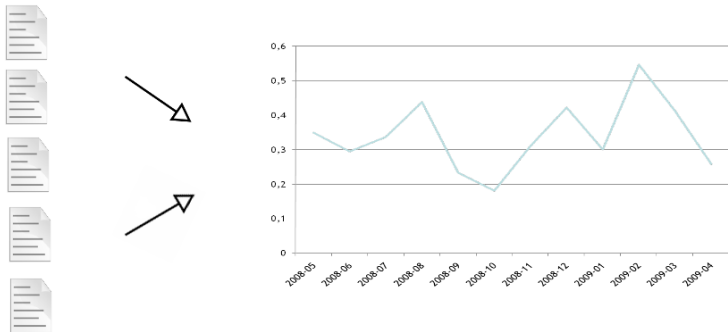
# UIMA in Blogoscopie project

Industrial integration workflow via a REST WebService



# UIMA in Blogoscopie project

From text collection to (graphical) informations in the same workflow



**Input** : All texts dealing with *Raymond Domenech* since 1 year

**Output** : Graphical evolution of positive/negative ratio  
for R. Domenech

# UIMA in Piithie project

- Objective : *Detection of text re-use*
  - Plagiarized texts
  - Tracking of a document's impact
- Industrial partners : Advestigo, Sinequa
- Tools needed for this NLP :
  - Morphological/Syntax annotator (Brill/Flemm),
  - Named Entities annotator (Nemesis)
  - Terminology extraction (Acabit)
  - Discursive markers analysis

# Overview

- 1 Introduction
- 2 UIMA : a solution for academic research
- 3 Using UIMA in projects : Blogoscopie and Piithie project
- 4 Conclusion : toward a french community around UIMA

# Toward a french community around UIMA

- 1st French Meeting around UIMA : July, 8th-9th in Nantes (RMLL/LSM) with academics and industrials
- Toward a portal to work collaboratively
  - ① Software repository (UIMA components, bug reports, patches, etc)
  - ② Experience exchanges (publications related with UIMA and NLP)
  - ③ Component and resource Developments for french
  - ④ Converging Type Systems for common annotations (morphosyntax, document, collection)
  - ⑤ <http://www.uima-fr.org>
- Facilitate common projects with other laboratories and industrials