# Large-scale Information Extraction from Neuroscientific Literature

Marco Antognini

Spring 2015

École Polytechnique
Fédérale de Lausanne

# Motivation

> The ultimate goal of the Blue Brain Project is to reverse engineer the mammalian brain.

- Interactions between brain regions?
- What about effects of specific cells?
- And connections with other organs?
- …and many many more questions.

The neuroscientific literature already holds many answers but…

# MOTIVATION

> The ultimate goal of the Blue Brain Project is to reverse engineer the mammalian brain.

- Interactions between brain regions?
- What about effects of specific cells?
- And connections with other organs?
- . . . and many many more questions.

The neuroscientific literature already holds many answers but. . .

$\Longrightarrow$. . . we need a tool to extract specific information from this colossal amount of text.

# UIMA & RUTA: IN BRIEF

# UIMA

- Unstructured Information Management Architecture
- Engines produce annotations (metadata)
- Engines are combined to form pipelines
- Most engines are written in Java
- Not limited to neuroscience topics

# UIMA – Sentence

Assuming we want to find all sentences in this text…

> Terminologies which lack semantic connectivity hamper the effective search in biomedical fact databases and document retrieval systems. We here focus on the integration of two such isolated resources, the term lists from the protein fact database UNIPROT and the indexing vocabulary MESH from the bibliographic database MEDLINE.

# UIMA – METADATA

The resulting annotations, presented in JSON format:

```
"DocumentAnnotation" : [
  { "begin" : 0,    "end" : 328,  "language" : "en" }
],
"Sentence" : [
  { "begin" : 0,    "end" : 135,
    "componentId" :
        "de.julielab.types.OpenNLPSentenceDetector" },
  { "begin" : 136,  "end" : 32
    "componentId" :
        "de.julielab.types.OpenNLPSentenceDetector" }
]
```

# BLUIMA

Bluima regroups UIMA engines, focusing on neuroscientific engines:

- Preprocessing (sentence, tokenizer, PoS, . . . )
- Linnaeus (species recognition)
- Oscar (chemistry)
- Brain regions & their relations
- Proteins
- Measures and units
- . . .

# RUTA

- Rule-based Text Annotation
- Scripting language
- Can integrate engine in script
- Makes it a bit easier to write UIMA pipeline

# RUTA – Simple Script

Basic RUTA script tagging dogs as Animal:

```
DECLARE Animal;
W{REGEXP("dog") -> MARK(Animal)};
```

# RUTA – Part of Speech Pipeline

Integration of UIMA engines:

```
ENGINE SentenceAnnotator;
ENGINE TokenAnnotator;
ENGINE PosTagAnnotator;

Document{-> EXEC(SentenceAnnotator)};
Document{-> EXEC(TokenAnnotator)};
Document{-> EXEC(PosTagAnnotator)};
```

# RUTA – Part of Speech Pipeline

Integration of UIMA engines:

```
ENGINE SentenceAnnotator;
ENGINE TokenAnnotator;
ENGINE PosTagAnnotator;

Document{-> EXEC(SentenceAnnotator)};
Document{-> EXEC(TokenAnnotator)};
Document{-> EXEC(PosTagAnnotator)};
```

BUT the engines' settings are stored in an unfriendly XML file...

# Issues

With both raw UIMA and RUTA scripts, some additional issues:

➤ Every user has to go through the hassle of installing UIMA/RUTA

➤ Not so trivial to run pipelines

➤ Manage external resources (install, update, remove, . . . )

➤ Manage versions of engines & pipelines

# Sherlok

# Sherlok – Text-Mining Service

In a few words:

- RESTful service: 1 server for many users
- Based on UIMA $\Longrightarrow$ existing engines compatible
- Based on RUTA $\Longrightarrow$ allowing powerful scripts
- Makes it easy to configure engines and pipelines
- Automatic resource management
- Versioning of pipelines

# Sherlok – Text-Mining Service

In a few words:

- RESTful service: 1 server for many users
- Based on UIMA $\Longrightarrow$ existing engines compatible (almost)
- Based on RUTA $\Longrightarrow$ allowing powerful scripts
- Makes it easy to configure engines and pipelines
- Automatic resource management
- Versioning of pipelines

Sometimes, the Java implementation needs some minor refactoring.

# SHERLOK – RESTFUL

What does it mean?

Basically, the HTTP protocol is all is needed to communicate with Sherlok.

$\Longrightarrow$ Can easily be embedded in any programming language or tool!

# Sherlok – RESTful

What does it mean?

Basically, the HTTP protocol is all is needed to communicate with Sherlok.

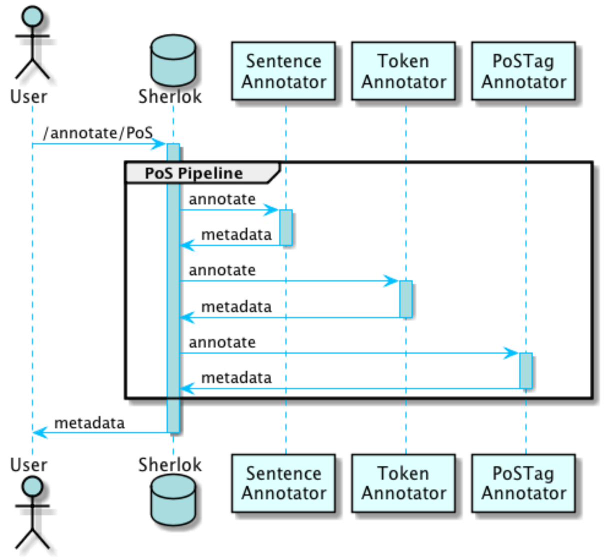
$\Longrightarrow$ Can easily be embedded in any programming language or tool!

---

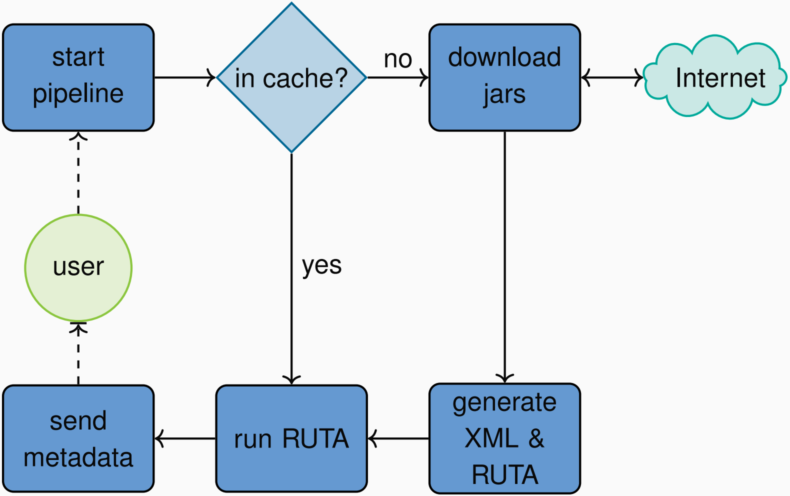
To annotate some text with the PoS pipeline:

```
GET /annotate/bluima.pos?text=<...>
```

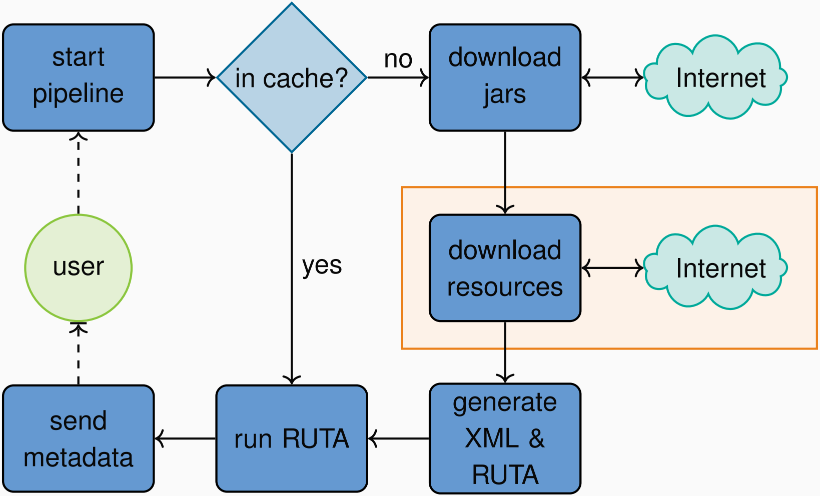
# Sherlok – Part of Speech Pipeline



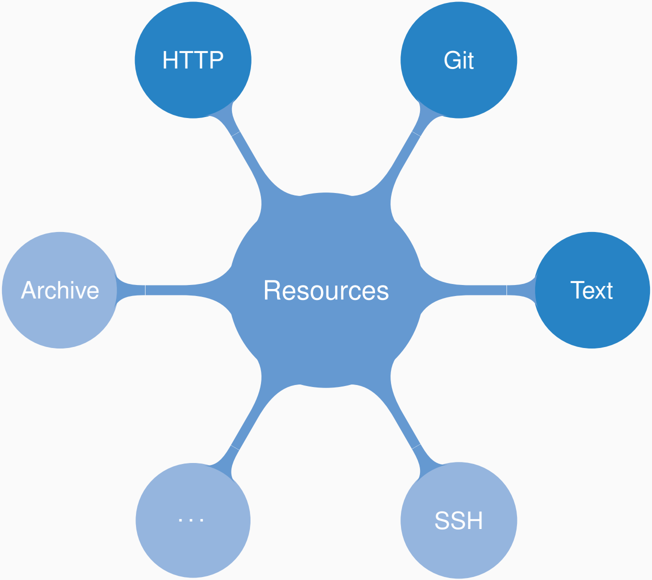
# Sherlok – Running a Pipeline



# Sherlok – Running a Pipeline



# Sherlok – Kinds of Remote Resource



# Sherlok – Bundle: Engine Configuration

```
{
  "name": "bluima.sentence", "version": "1.0.1",
  "dependencies": [
    { "value": "ch.epfl.bbp.nlp:bluima_opennlp:1.0.1" }
  ], "config": {
    "bluima": {
      "type": "git", "ref": "master",
      "url":
        "https://github.com/BlueBrain/bluima_resources.git" }
  }, "engines": [ {
      "name": "SentenceAnnotator",
      "class": "ch.epfl.bbp.uima.ae.SentenceAnnotator",
      "parameters": {
        "modelFile":
          "$bluima/opennlp/sentence/SentDetectGenia.bin.gz"
      }
  } ]
}
```

# Sherlok – Pipeline Configuration

```
{
  "name": "countries", "version": "1",
  "description": "Example that annotates countries",
  "config": {
    "countries": {
      "type": "http", "mode": "ruta",
      "url": "https://example.com/countries.txt"
    }
  }, "script": [
    "WORDLIST CountriesList = ’$countries’;",
    "DECLARE Country;",
    "Document{-> MARKFAST(Country, CountriesList)};"
  ]
}
```

# DEMO!

# Thank you!

Questions?

Check it out

http://sherlok.io