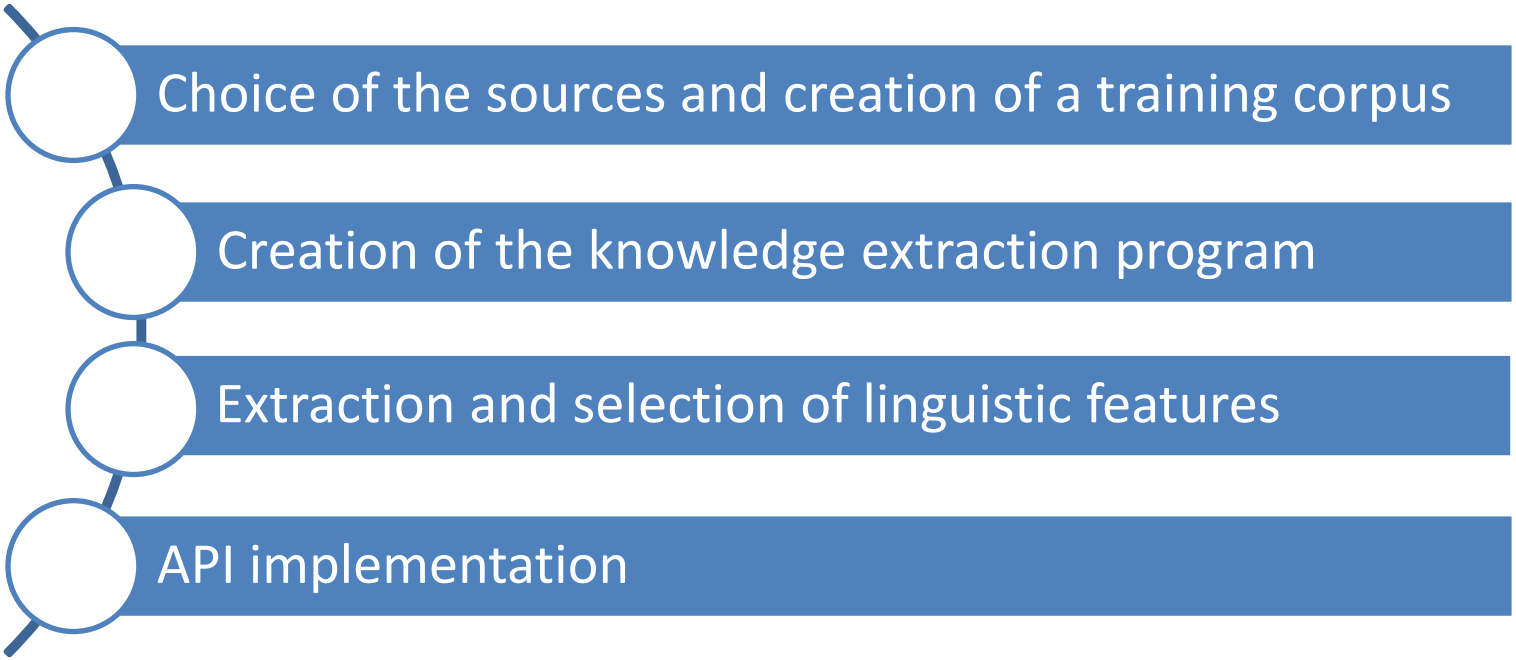
Automatic extraction of IDM-related information in scientific articles and online science news websites

Laboratoire CSIP

TRIZ Future Conference 2018 - Strasbourg

NÉDEY Oriane, SOUILI Achille, CAVALLUCCI Denis

Project outline



Choice of the sources and creation of a training corpus

Two training corpora in order to :

- Improve the performances for the extraction
- Assess the performances of the extraction program

A database :

- 15500+ sciences news articles from 2016-2018
- 550+ research articles

Science news articles

- Machine Design
- New Atlas
- Phys.org
- Research & Development
- Science Daily
- Science News
- Science News for Students

Research articles

- Accounts of Chemical Research
- Annual Review of Condensed Matter Physics
- Chemistry of Materials
- Proceedings of the National Academy of Sciences of the USA

API Creation

Preprocessing

- News processing from URLs
- Articles cleaning – from HTML to JSON format

Extraction : *Problems, Partial solutions, Parameters*

- Structure : analogy with another knowledge extraction program, working with patents
- Differenciated extraction (type of article, sections…)
- Extraction with linguistic features

Features Extraction and selection

A significant **drawback** associated with this approach is that a number of different kinds of bacteria have carbohydrates.

- Use of previous sets of features, working with patents
- Assessment, extraction and final selection of features : possible with annotated corpora
- Examples of new linguistic features added to our sets :
  - by (solutions partielles)
  - expensive (problèmes)

Perspectives

Improving performances

- Build a bigger training corpus
- Divide the articles into more precise sentence-like units
- Modify the program regarding parameters extraction

Fonctionalities

- Implement the API in a web application :
  - Search for articles in a database or from a URL
  - Automated creation of a problem graph
- Add new possible article sources
- Distinguish Action Parameters (AP) from Evaluation Parameters (EP)
- Reformulate the extracted sentence-like units
- Add the automated extraction of values (and opposite values)