OpenMinted

1. Deliverable T.4: Final Report

2. Project: UIMA FreeLing integration

3. Contact Person

Joan Codina
TALN group, Universitat Pompeu Fabra
joan.codina@upf.edu
+34 93 542 2241
C. Tànger, 122-140
08018 Barcelona

4. Final Report

The component is shared as a UIMA component (available in GitHub) and integrated together with FreeLing in a Docker image (also available in hub.docker.com)

Reasons to do a Docker component

The reason to share the component as a Docker image was for the difficulties to run FreeLing as a java component. FreeLing is written in C++ and includes several data folders. All this information is integrated in the docker, so the user doesn't have to worry about the FreeLing installation itself.

Highlights and key success factors

The integration of the FreeLing parser to the UIMA environment has been carried out. The component takes as input an XMI containing the text. The text is processed and the output includes the tokens, lemmas, POS and dependencies adapted to the DKPro-Core tagset.

The integration involved the analysis and definition of the correspondence tables for the tagsets provided by FreeLing for each language and annotation level to the DKPro-Core tagset.

The component is delivered as a docker that includes the FreeLing installation. As FreeLing is written in C++ while UIMA in Java, the deployment of a UIMA component that includes FreeLing using Maven would be too complex. The proposed solution is the integration for OpenMinTeD with a docker. The UIMA component is also published as an open source project in GitHub under a AGPL license. Therefore, this component also allows the usage of Freeling as UIMA component in any UIMA pipeline so that it can be easily combined with other OpenMinTed components.

The Freeling component provides basic language analysis functionalities (tokenization, lemmatization, Pos Tagging and dependency parsers.) for the variety of languages that Freeling includes (English, Spanish, Portuguese, Italian, French, German, Russian, Catalan, Galician, Croatian, Slovene) within the context of OpenMinTeD. The specific usage scenario for this prototype concerns scientific publications in non-English languages. This means that the considerable amount of scientific production in non-English languages will be incorporated thanks to the implementation of this component.

Unexpected events

Even that everything went more or less smoothly the where some problems:

- With the different versions of the software and compatibility between them. Building the FreeLing docker was more complicated than expected.
- We found some errors in FreeLing that we reported and the people of Freeling solved very quickly but this means that we had to move to the development version of Freeling
- We had some difficulties when integrating our component to the OpenMinted platform, the documentation was no clear to us. But we got the right support from the lists.

Balance of the amount of work to adapt and integrate Freeling into OMTD

- We found the task a bit bigger than expected due to the unexpected events: even that we already thought that some problems could arise, debugging into the docker was complex.

Justification of the work done

- We think that effort done has produced a reusable component, so it is justified, the integration of Freeling into the platform allows the processing of text in languages that otherwise are discriminated.

Lessons learned from this project

- For us has been important to view and understand the way to organize and share applications and resources between researchers.

What we would have done differently and/or our recommendations for improvement

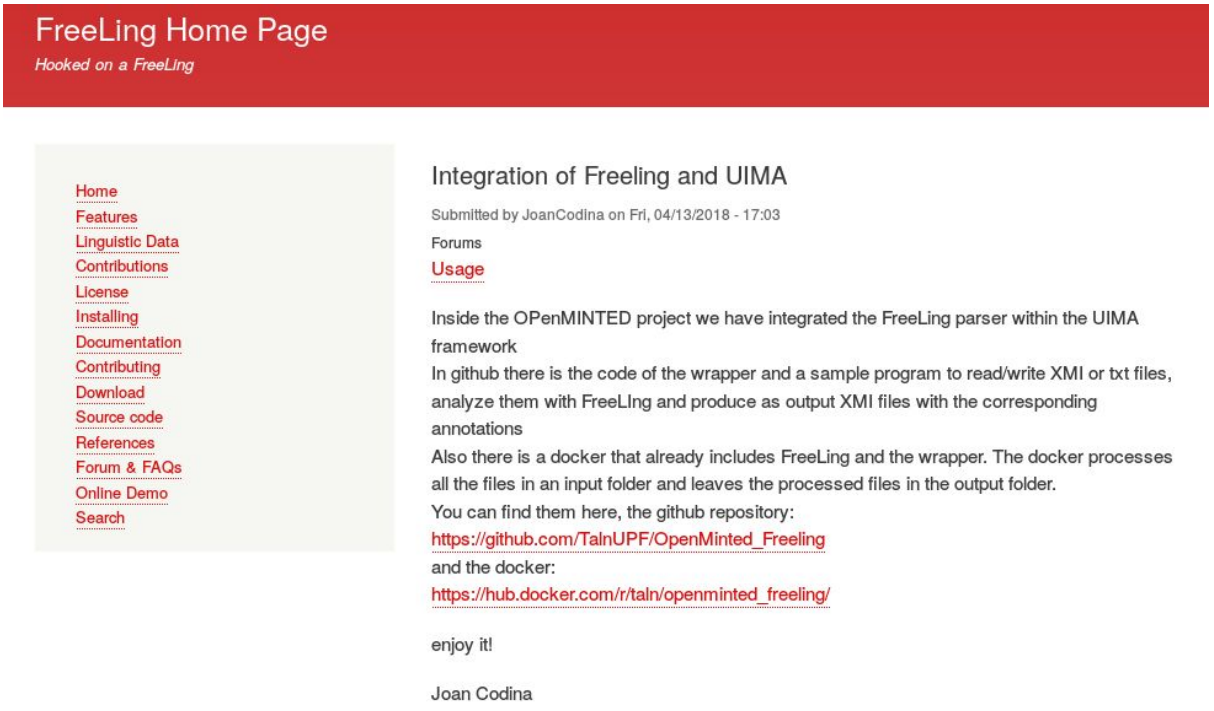
- Timing has been very short.
- It might have been interesting to have a tutorial on how to integrate components, and to have a single skeleton of each kind of component.

5 Dissemination Report

The component has been uploaded in GitHub.com, and hub.docker.com

We have also communicated to the FreeLing community the existence of the component

http://nlp.lsi.upc.edu/freeling/node/645



A blog post entry has been sent to Martine Oudenhoven.

Integration of Freeling into the OpenMinTeD platform

An implementation of Freeling as a docker service and shared as a UIMA component has been carried out within the context of OpenMinTeD. The Freeling component provides basic language analysis functionalities (tokenization, lemmatization, Pos Tagging and dependency parsers.) for the variety of languages that Freeling includes (English, Spanish, Portuguese, Italian, French, German, Russian, Catalan, Galician, Croatian, Slovene). The specific usage scenario for this component concerns scientific publications in non-English languages. This means that the considerable amount of scientific production in non-English languages will be incorporated thanks to the implementation of this component. The component, which has been shared as a Docker and the code released in Github with a GPL license, can be found in the following links:

https://hub.docker.com/r/taln/openminted_freeling/

https://github.com/TalnUPF/OpenMinted_Freeling

Adapt course:

The course "OpenMINTED: FreeLing and BabelNet Components". Has been uploaded to the adapt platform: http://courses.fosteropenscience.eu