ROMANIAN PROCESSING CHAINS IN METANET4U

PISTOL IONUȚ CRISTIAN

"Alexandru Ioan Cuza" University, Faculty of Computer Science, Iași – Romania

ipistol@info.uaic.ro

Abstract

This paper's main contribution is to describe the completed and planned development of processing resources at UAIC[1] as part of the work done for the METANET4U[2] research project. Significant for the project is the development of processing resources as UIMA[3] components integrated in the U-Compare[4] system, which offers significant advantages in terms of workflow development and evaluation.

1. Introduction

The main goal of the METANET4U project is to collect language resources for seven European languages and distribute them using a platform called METASHARE. The contributed resources can have many forms, from annotated corpora to complex processing systems and from open-source tools to pay-per-use web services.

Complex NLP applications such as information extraction systems comprise several separate tools such as tokenizers, part-of-speech taggers, named entity recognizers, etc. Whilst these tools may be developed for the purposes of a particular application, it is desirable if they can be re-used in other applications. This is because the same basic processing steps are often common for a number of different NLP applications. As part of METANET4U, work has been carried out on the WPS Work package, whose main goal is to show if and how processing tools originating in various sources and usable for various languages can be combined to build complex processing workflows.

Supporting this goal, U-Compare (Kano et. al, 2011) is a workflow management system based on UIMA (Ferrucci and Lally, 2004), a well know NLP meta-system allowing users to contribute processing tools and use them together with other integrated resources to perform various processing tasks.

As part of UAIC's contribution to METANET4U (and the WPS Work package), we selected 18 processing tools developed at UAIC to be contributed to METASHARE (14 of which will be integrated in UIMA and U-Compare). This paper describes part of this work, next section making a short overview of UIMA and U-Compare, as well as the effort required to integrate a new tool. UAIC tools and the current state of the integration is described in section three. Integration issues and future considerations are discussed in the last section of this paper.

[1] http://www.uaic.ro/uaic/bin/view/Main/?language=en
[2] http://metanet4u.eu/
[3] http://uima.apache.org/
[4] http://u-compare.org/

2. UIMA/U-Compare integration

UIMA (Unstructured Information Management Architecture) is the result of an IBM development project (completed in 2002) aiming to develop an "industrial-strength, scalable and extensible platform for creating, integrating and deploying unstructured information management solutions from combinations of semantic analysis and search components." (Ferrucci and Lally, 2004). It was designed for maximum performance and scalability, intended to serve as a linguistic annotation black-box used to add whatever linguistic information was available to any type of data. By "unstructured information" IBM meant all types of available electronic resources, as a whole, without a common structure. UIMA's goal was to process all this data and add linguistic information and structure to it, thus significantly improving classification, advanced search and data transfers.

U-Compare (Kano et al., 2011) has been developed by the University of Tokyo, the National Centre for Text Mining (NaCTeM) at the University of Manchester and the University of Colorado, with the goals to support construction of NLP applications from reusable resources and to allow easy evaluation of applications against gold-standard annotated data. U-Compare is based UIMA and inherits UIMA's description of annotations as Types (basically each annotation is an instance of a particular UIMAType class, offering access to read existing elements and writ new ones observing a specified Type specification). This has the benefit of guaranteeing interoperability between components using the same Types as input/output, but has the significant drawback of requiring users to adapt their tools to access annotated data not directly from an external resource but internally, using access methods available for that particular Type.

Before METANET4U, U-Compare included a set of over 30 integrated processing resources, most of them available for English. For those tools, U-Compare offered means to combine them in various workflows using a graphical interface, which serves as a repository of available resources and allows users to check whether the components added in sequence to a workflow actually match input/output formats (indicated as specific U-Compare Types part of the U-Compare Type System.

Since one of the main developers of U-Compare (University of Manchester) is also part of METANET4U and the leader of WPS, U-Compare has been adapted to the conditions and issues raised so far during the project, particularly in terms of handling multilingual components and workflows created.

3. UAIC WPS current status

The first stage involving UAIC required us to select tools we can contribute to METASHARE. We selected 18, all developed at UAIC (and all available for free, either as open source or web service). Of them, 14 were selected for integration in UIMA/U-Compare. We kept the tools relevant in multilingual contexts, performing tasks relevant for other languages if the required resources are provided. Table 1 below shows the 15 UAIC tools to be integrated, together with the selected U-Compare Type System input and output format.

Table 1: UAIC tools in WPS

Tool name	Input	Output	Observations
Splitter	org.u_compare.shared.document.Text	org.u_compare.shared.document.Segment (new type added by UAIC)	Splits to discourse units
Tokenizer	org.u_compare.shared.document.Text	org.u_compare.shared.syntactic.Token	
Lemmatizer	org.u_compare.shared.syntactic.POSToken (or Text)	org.u_compare.shared.syntactic.RichToken	Two versions with different input type
FDG-parser	org.u_compare.shared.syntactic.POSToken	org.u_compare.shared.syntactic.Dependency	
NP-chunker	org.u_compare.shared.syntactic.POSToken	org.u_compare.shared.syntactic.Chunk	
RARE	org.u_compare.shared.syntactic.Chunk	org.u_compare.shared.semantic.CoreferenceAnnotation	Performs anaphora resolution
Discourse Parser	org.u_compare.shared.semantic.CoreferenceAnnotation	org.u_compare.shared.semantic.DiscourseTree (new type added by UAIC)	
SRL	org.u_compare.shared.syntactic.RichToken	.u_compare.shared.semantic.SemanticClassAnnotation	Performs semantic role labeling
Summarizer	org.u_compare.shared.document.Text	org.u_compare.shared.document.Text	Output is a different UIMA view of the same document
OntologyBuilder	org.u_compare.shared.syntactic.RichToken	org.u_compare.shared.syntactic.OWL (new type added by UAIC)	Builds an ontology from keywords and definitions
QA	org.u_compare.shared.document.Text	org.u_compare.shared.document.Text	Output is the answer to the input questions
TE	org.u_compare.shared.document.Text	org.u_compare.shared.document.Text	Checks entailment between two input fragments
OccurenceFinder	Any	Keeps original format	Finds occurrences of an annotation pattern
Categorizer	org.u_compare.shared.document.Text	org.u_compare.shared.document.Category (new type added by UAIC)	Labels text with general semantic categories

Using the above tools as well as those contributed by other project members, a set of 26 multilingual workflows were designed to be implemented by July 2012 (Branco et al., 2011). 22 of the 26 workflows involve components developed by UAIC. An example of such a workflow can be seen in figure 1.

Of the tools in Table 1, the first 3 are already integrated in UIMA/U-Compare and were use to build and test 4 workflows (two of them using also components developed by RACAI, the other METANET4U partner from Romania). An example of one of these workflows, as it appears in U-Compare's interface, can be seen in figure 2. This particular workflow uses plain text as input and produces tokenized, POS-tagged and lemmatized output.

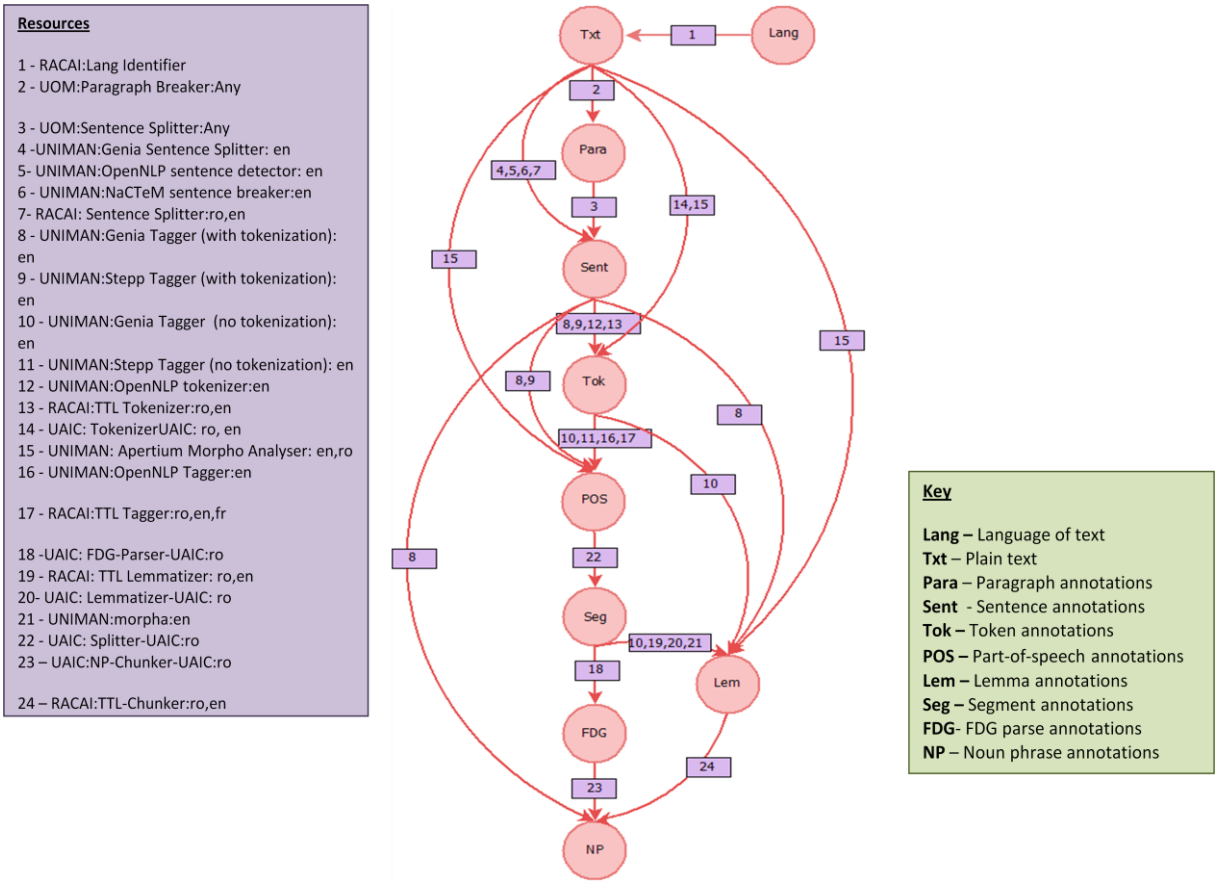

Figure 1: An example of a multilingual WPS workflow (adapted from (Ananiadou et. Al. 2011))

The integration faced some difficulties, requiring adaptation of both U-Compare and individual tools to satisfy requirements apparent only after the integration begun, such as issues with exporting components and workflows within U-Compare and accessing external resources (dictionaries, language models) required by some components. These difficulties were largely overcome, and the integration of the next set of UAIC components is under way.

A significant change required for most UAIC or other partner's tools is to adapt to a common standard way of reading and writing annotations, which usually involves changing the current implementations. This is usually manageable for endogenous tools, where some of the original developers are available to make changes.

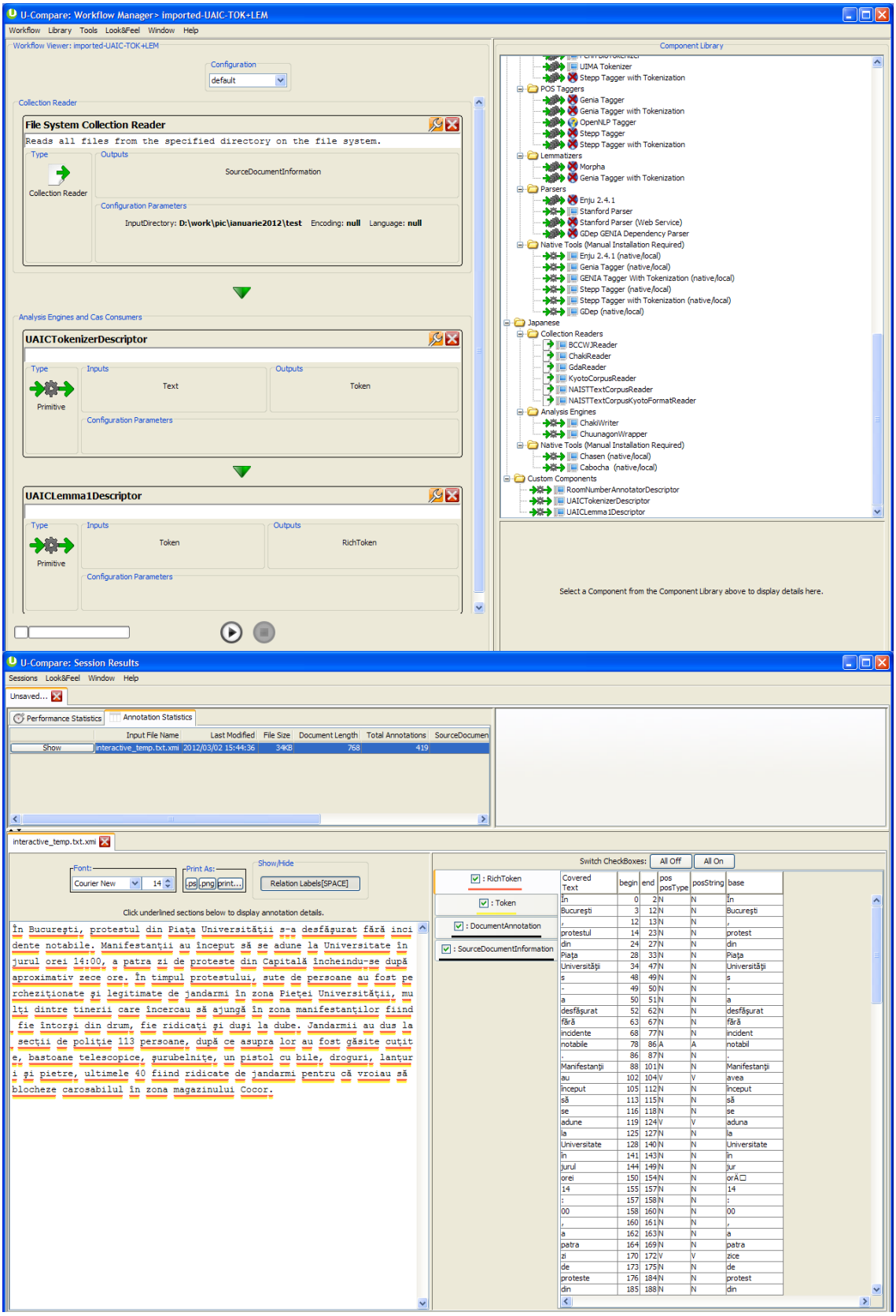

Figure 2: A workflow using UAIC components (above) and the results produced in U-Compare for a short text (below)

4. Conclusions

The benefits of collecting NLP resources from multiple developers and many languages and showing how they can be combined and compared is significant, both for application developers and the uninformed user of NLP techniques. The developer benefits knowledge of other similar tools, independent comparison of the results and guaranteed compatibility with relevant other components. The uninformed user can select available workflows without knowing their internal architecture, and can be assured that the components selected are compatible and working with the efficiency provided by the UIMA integration.

The benefits for the Romanian language are most of all of visibility, the large set of language processing components contributed by the Romanian partners (second largest in METANET4U, after English) proves again that Romanian is on the leading edge of NLP development.

What projects like METANET4U prove is that standardization brings significant advantages only if it involves large sets of developers and allows for some flexibility. The work carried out so far showed that open source components, web services and proprietary software can work together seamlessly if a minor standardization effort is made by motivated partners.

Acknowledgements. The work described in this paper was partially supported by the METANET4U EC CIP project #27089.

References

Branco, A., Trancoso, I., Ananiadou, S., Thompson, P., McNaught, J., Cristea, D., Tufis, D., Rosner ,M., Moreno, A., Bel, N. (2011). Specification of pilot services and applications. *Document METANET4U-2011-D2.2*, EC CIP project #270893, available on http://metanet4u.eu/.

Ferrucci, D., Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment, *Natural Language Engineering 10*, No. 3-4, 327-348.

Kano, Y., Miwa, M., Cohen, K., Hunter, L., Ananiadou, S., Tsujii, J. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55: 3, 11:1-11:10.