Integration and deployment of Unitex-based applications in a lightweight web services architecture

Cristian Martinez [*], **Amina Marie** [**]

[*] *Paris-Est University, Gaspard-Monge Computer Science Laboratory (LIGM)*
Model and Algorithms (MoA) research team. 77420 Champs-sur-Marne, France.
cristian.martinez@univ-paris-est.fr
[**] *AMABIS, 92340 Bourg-la-Reine, France.*
amina.marie@amabis.fr

Abstract: Unitex is an open source, cross-platform and multilingual text corpus processing suite. Unitex tools are designed to be run on a textual corpus, performing several natural language processing (NLP) tasks using linguistic resources. A web service is a self-contained, self-described application entity that is deployed, published and invoked over the network using open protocols. Web services facilitate applications to interoperate in a loosely coupled way allowing for the creation and integration of other services. While NLP engines are the core of the workflow process for some real-world applications, over the past few years, web services technologies have become the de-facto standard for exposing services and implementing business collaborations. Consequently, the need to integrate the Unitex environment into the web services model becomes more and more present. In this work, we present a lightweight web services architecture aiming at the integration and deployment of Unitex-based applications. The proposed approach addresses issues raised by the growing use of Unitex in heavy processing projects for both scientific and industrial purposes, as well as it delivers a highly available, reliable and easy to scale solution to expose Unitex projects as ReST (Representational state transfer) web services. The main components of our architecture are: a dynamic load balancer, a message broker, a set of task queues, a pool of distributed workers, and a Unitex abstraction layer (UAL) library. Web services public endpoints are exposed via APIs that use ReST-Like semantics and binary JSON (JavaScript Object Notation) as a data transport format. When a service consumer sends a request message, a load balancer forwards it into a server broker, then the broker generates a unique response channel identification, transforms the message to an alternative job representation and places it in a named queue. The jobs are then executed concurrently on a pool of distributed workers that run multiple, identical service instances which are associated with the name of a queue. Workers are service-oriented components which are in charge of job processing using the UAL interface, deleting them once the procedure is terminated and writes back a response message using the reply channel. The UAL library mainly provides a simple interface on top of the Unitex native API, using a pattern that enable to describe a workflow only once and then execute it many times. The workflow gives the possibility to specify inputs, outputs, resources and the sequence according to which the system and Unitex tasks have to be performed. In order to evaluate the proposed approach, we have built a proof-of-concept service for postal address processing, i.e. providing the ability to identify, validate and enhance the components of an unparsed address string. The NLP analysis engine consists of a set of Unitex local grammars coupled to electronic dictionaries describing both the structure and the constituent elements of an address : given name, surname, street number, street name, postal code, locality, country, etc. We discuss about the high availability and small footprint of this implementation, as well as other strengths and weaknesses. Finally, we conclude with further perspectives to extend the proposed architecture in order to include supervision, administration, accounting, logging and authentication capabilities.

Keywords: distributed systems; natural language processing; rest; unitex; web services.