Web Service integration platform for Polish linguistic resources

Maciej Ogrodniczuk and Michał Lenart

Institute of Computer Science, Polish Academy of Sciences

INTRODUCTION

The Multiservice is a webservice providing a common interface and coherent linguistic annotation of Polish texts basing on individual offline tools.

- one service for chaining execution of linguistic tools
- processing triggered by requests sent to the webservice,
- requests enqueued and handled in asynchronous manner,
- single request specifies a list of processing chain parts:
  - operation type (e.g. shallow parsing),
  - requested tool name (e.g. parser name),
  - a map of tool-specific properties (e.g. whether a tag filter should be used before running the parser).

Advantages

As with any webservice:

- no dependencies to download, no configuration to perform,
- access via a portable and simple SOAP-based API,

plus more:

- an easy-to-use Python script,
- asynchronous processing valid for large amounts of texts,
- possibility to distribute the processing across multiple machines.

USAGE

General usage scheme is following:

1. a user sends a request containing processing chain settings,
2. the service returns a unique token,
3. the user keeps asking about request status util it is DONE or FAILED ...
4. ... and fetches the result (or an error message).

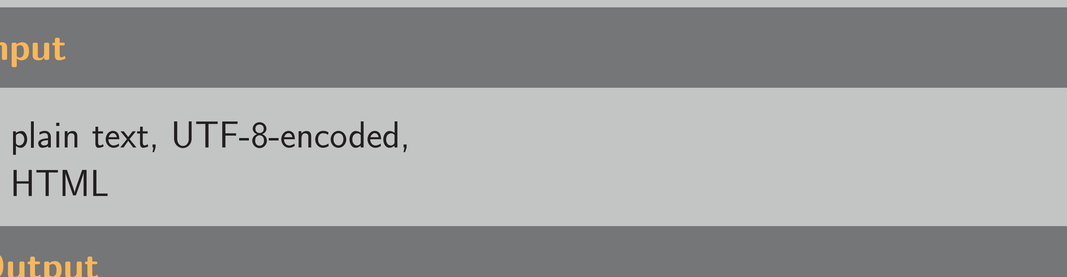
Input

- plain text, UTF-8-encoded,
- HTML

Output

- a "packaged" version of XML TEI P5-based format of the National Corpus of Polish (see `http://nlp.ipipan.waw.pl/TEI4NKJP/`),
- linguistic features preserved using TEI-embedded feature structure formalism,

ARCHITECTURE



The Webserver:

1. receives the request and writes it to the database with PENDING status,
2. notifies request processing daemons about arriving request.

A daemon:

1. gets the request from the database,
2. marks it as IN PROGRESS,
3. calls subsequent underlying language services to perform it,
4. once all parts of the request are processed, writes the result to the database and marks it as DONE (or FAILED – and then returns an error message).

WEB INTERFACE

- Available at `http://chopin.ipipan.waw.pl/multiservice/`,
- user can construct a processing chain using simple, clickable interface,
- results can be viewed either as raw XML output or visualized in human-readable way

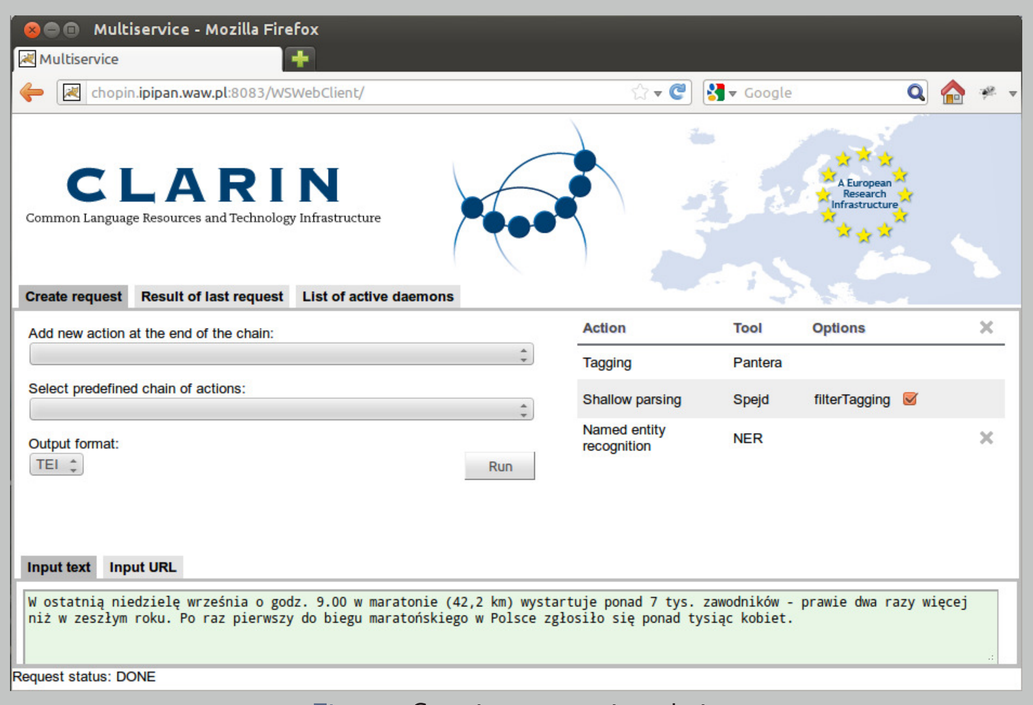


Figure: Creating processing chain

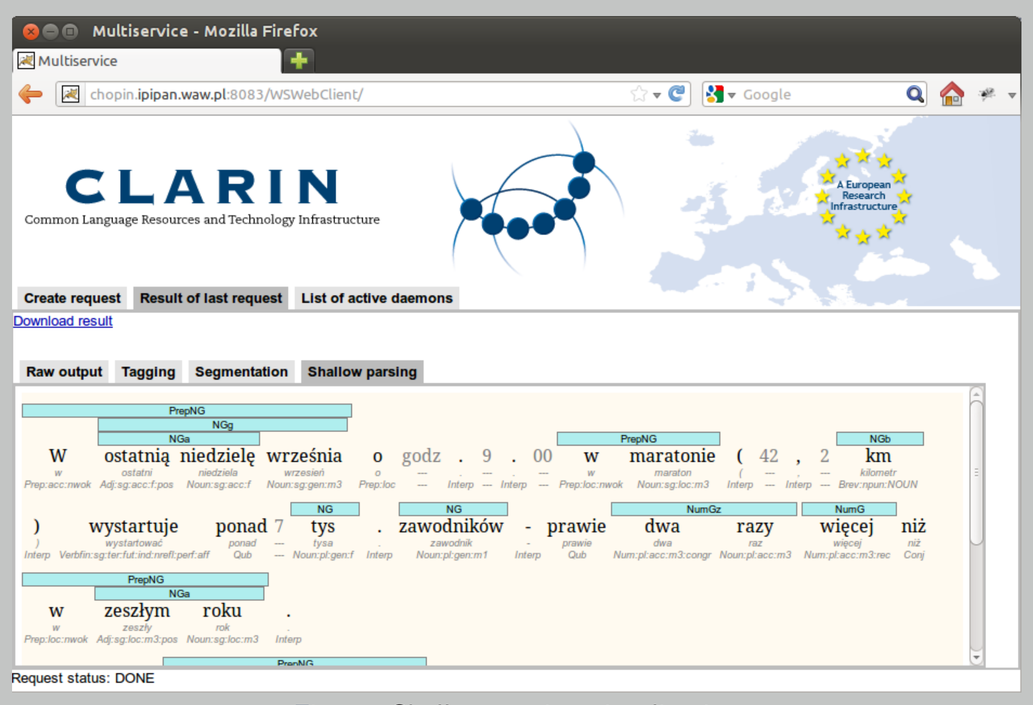


Figure: Shallow parsing visualization

NEW IMPLEMENTATION

- Apache Thrift is used for data exchange and RPC (instead of XML or plain text)
- It is a mature project created and used by Facebook and many other software companies.
- Does remote calls across various languages (C++, Python, Java, Ruby, ...) using fast binary protocol,
- provides simple API for creating a TCP server for RPC calls

Main benefits

- Much less overhead (result XML document generated only at the very end of request processing),
- easier and more coherent integration of new language tools (provided they are written in a language supported by Thrift)

REFERENCE

- Maciej Ogrodniczuk and Michał Lenart. *Web Service integration platform for Polish linguistic resources*. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, pages 1164–1168, Istanbul, Turkey, 2012. ELRA.

michal.lenart@gmail.com, maciej.ogrodniczuk@ipipan.waw.pl