U-Compare Tokenisation Service

1. BASIC INFORMATION

Service name

U-Compare Tokenization service

Overview and purpose of the tool

This is a web service that identifies sentences and tokens in Portuguese text.

A short description of the algorithm

This web service is based on a UIMA-based workflow, created using the U-Compare text mining system[1]. The workflow was exported from U-Compare as a web service using the built in functionality (Kontonastsios et al., In Press). The workflow was created as part of the work to increase the number of interoperable tools operating on different European languages (Ananiadou et al, 2011).

The workflow consists of the following UIMA tools, in the specified order:

1) Freeling[2] sentence splitter web service[3] (service provided by the PANACEA project[4])
2) LX-Tokenizer[5] (web service provided by the University of Lisbon)

2. TECHNICAL INFORMATION

Software dependencies and system requirements

This is a web service that can be run from a browser or accessed programmatically. The only basic requirement is an internet connection.

Installation

There is no installation. The web service can be accessed at the following URL:

http://nactem001.mib.man.ac.uk:8080/UCompareWebServices/Tokenisation_Freeling_LX

[1] http://nactem.ac.uk/ucompare/

[2] http://nlp.lsi.upc.edu/freeling/
[3] http://registry.elda.org/
[4] http://www.panacea-lr.eu/
[5] http://lxcenter.di.fc.ul.pt/services/online_suite/en/lx-suite.html

The web form available at this URL is shown in Figure 1

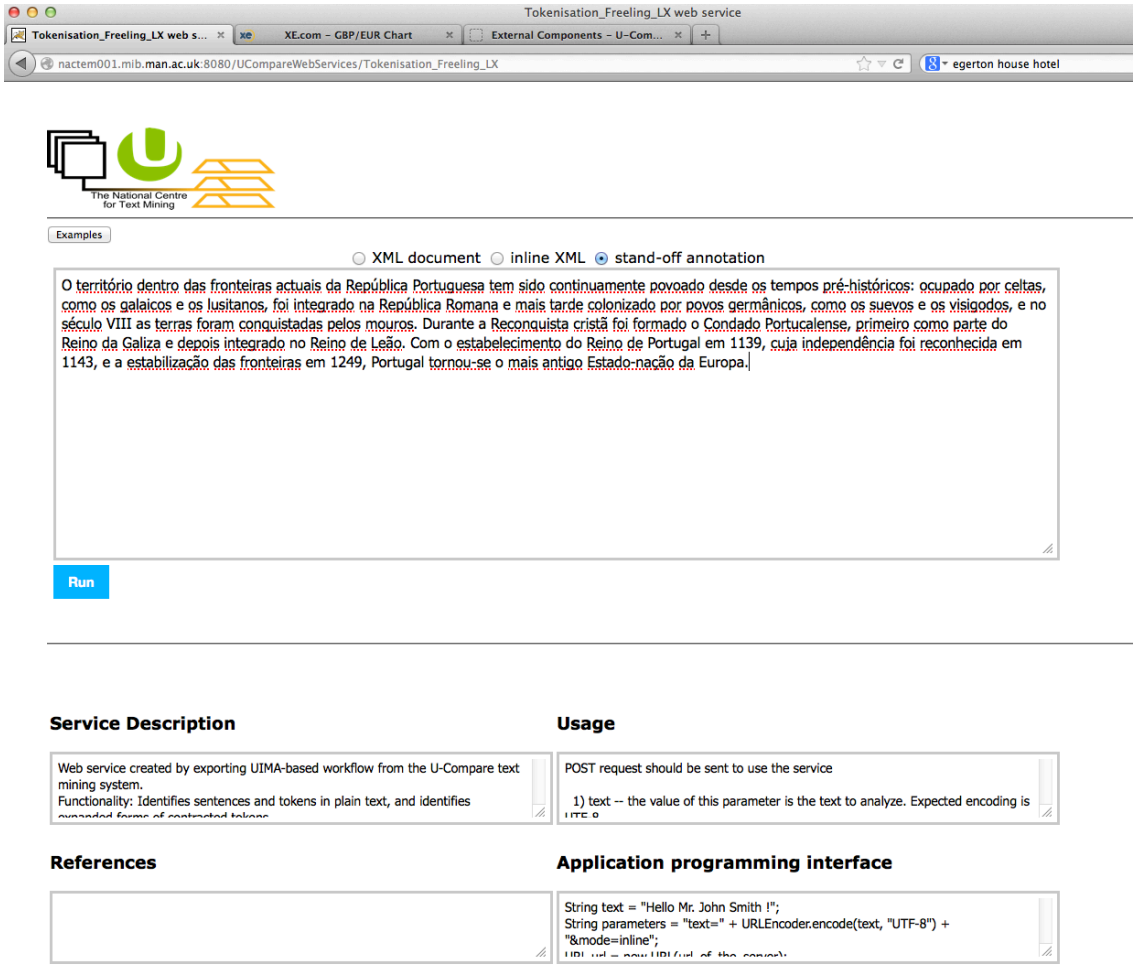


Figure 1: Web form for the U-Compare tokenisation service

Execution instructions

The web service can be executed by typing or pasting text into the online form and clicking on the "Run" button.

Alternatively, the web service can executed from within program code, as explained in the "Usage" and "Application programming interface" boxes of the web form.

A POST request should be used to call the service. The following parameters may be used in the request:

- **text** - the value of this parameter is the text to analyze. Expected encoding is UTF-8. This parameter is obligatory.
- **lang** - This parameter sets the language of the text. If this parameter is not provided, then the value"en" will be used

- **mode** - This parameter sets the format of the annotated information returned by the service. If this parameter is not set, XML output will be produced. The two possible types of output are as follows:
    - **inline** – annotations are encoded as inline XML.
    - **xml** – results are output as an XML document containing the annotations added

The following code example shows how the web service can be called from Java code:

```
//Set the input text
String text = " <Text_to_be_analysed>";
//Set the parameter string
String parameters = "text=" + URLEncoder.encode(text,
"UTF-8") + "&mode=inline";
//Create the URL connection
URL url = new
URL("http://nactem001.mib.man.ac.uk:8080/UCompareWebServi
ces/Tokenisation_Freeling_LX");
URLConnection connection = url.openConnection();
connection.setDoOutput(true);
//Create Output stream
OutputStreamWriter writer = new
OutputStreamWriter(connection.getOutputStream());
//write parameters to output stream
writer.write(parameters);
writer.flush();

//Read the results returned by the service
BufferedReader reader = new BufferedReader(new
InputStreamReader(connection.getInputStream(), "UTF-8"));
String line;
while ((line = reader.readLine()) != null) {
    System.out.println(line);
```

}

Input/Output data formats

Input data formats

The input is plain text, UTF-encoded.

Output data format

If the service is run from the web interface, then the output is visualized in the interface using colored highlights in the text to show the individual annotations, and one or more tables of information below, each corresponding to a particular type of annotation.

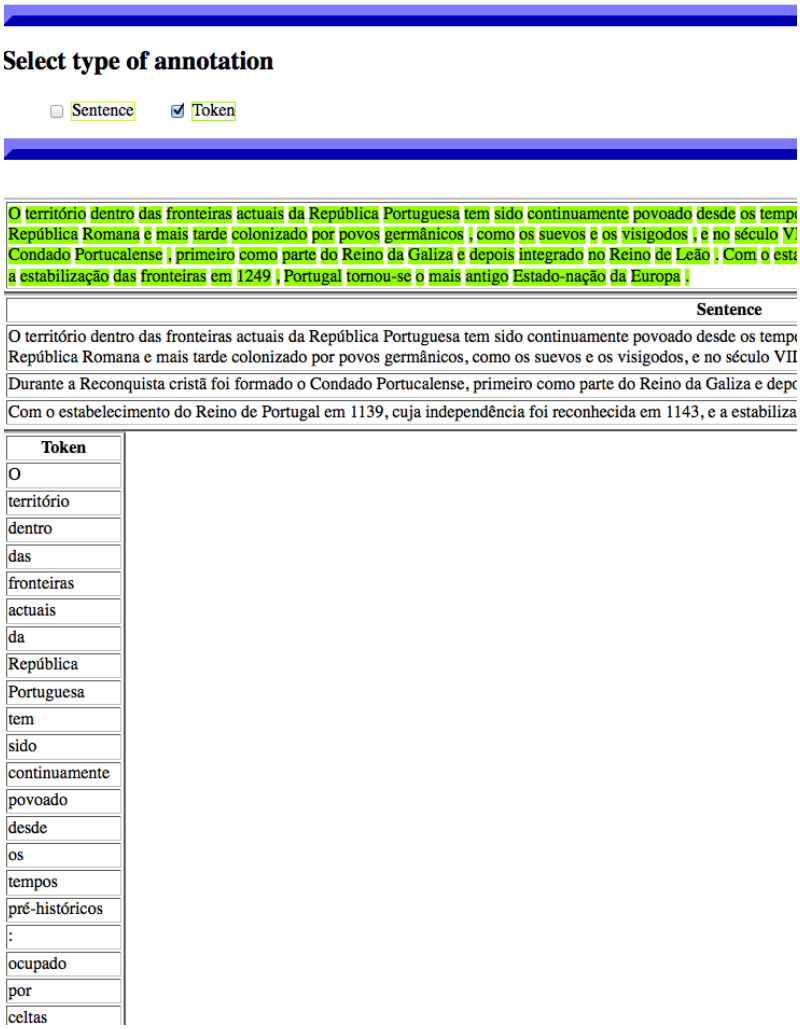
If the service is run programmatically, then the output is provided in XML format. See section 3 for an example.

Integration with external tools

The API allows the functionality of the web service to be embedded in any application.

3. CONTENT INFORMATION

Using the web interface, the output of the service is visualised as shown in Figure 2.



Select type of annotation

Sentence Token

O território dentro das fronteiras actuais da República Portuguesa tem sido continuamente povoado desde os tempo
República Romana e mais tarde colonizado por povos germânicos , como os suevos e os visigodos , e no século V
Condado Portucalense , primeiro como parte do Reino da Galiza e depois integrado no Reino de Leão . Com o esta
a estabilização das fronteiras em 1249 , Portugal tornou-se o mais antigo Estado-nação da Europa .

Sentence
O território dentro das fronteiras actuais da República Portuguesa tem sido continuamente povoado desde os tempo República Romana e mais tarde colonizado por povos germânicos, como os suevos e os visigodos, e no século VII
Durante a Reconquista cristã foi formado o Condado Portucalense, primeiro como parte do Reino da Galiza e depo
Com o estabelecimento do Reino de Portugal em 1139, cuja independência foi reconhecida em 1143, e a estabiliza

Token
O
território
dentro
das
fronteiras
actuais
da
República
Portuguesa
tem
sido
continuamente
povoado
desde
os
tempos
pré-históricos
:
ocupado
por
celtas

Figure 2: Visualisation of tokenisation results

In Figure 2, the top of the screen has check boxes corresponding to each type of annotation produced by the workflow – in this case “Sentence” and “Token” annotations. Checking one or more of the boxes will cause the annotations to become highlighted in the view of the text below. In figure 2, only “Token” annotations are highlighted.

Below the text, the different types of annotations added by the workflow are shown in tabular format, with each type of annotation in a separate table. In Figure 2, it can be seen that there are two tables, one for “Sentence” annotations and one for “Token” annotations. For each annotation type, the text covered by each individual annotation in shown in a row of the table.

An example of the XML output format, which is more suited to programmatic use, is shown in Figure 3. In the XML, the start and end offsets of each annotation in the text are encoded in the “begin” and “end” attributes.

```
– <result>
    <Token begin="0" end="2">În</Token>
    <Token begin="3" end="9">ajunul</Token>
    <Token begin="10" end="15">celui</Token>
    <Token begin="16" end="18">de</Token>
    <Token begin="18" end="19">-</Token>
    <Token begin="19" end="21">al</Token>
    <Token begin="22" end="28">Doilea</Token>
    <Token begin="29" end="35">Război</Token>
    <Token begin="36" end="43">Mondial</Token>
    <Token begin="44" end="45">(</Token>
    <Token begin="45" end="49">1940</Token>
    <Token begin="49" end="50">)</Token>
    <Token begin="50" end="51">,</Token>
    <Token begin="52" end="59">România</Token>
    <Token begin="60" end="64">Mare</Token>
    <Token begin="65" end="66">(</Token>
    <Token begin="66" end="75">Întregită</Token>
    <Token begin="75" end="76">)</Token>
    <Token begin="76" end="77">,</Token>
    <Token begin="78" end="81">sub</Token>
    <Token begin="82" end="91">presiunea</Token>
    <Token begin="92" end="101">Germaniei</Token>
    <Token begin="102" end="109">naziste</Token>
    <Token begin="110" end="117">condusă</Token>
    <Token begin="118" end="120">de</Token>
    <Token begin="121" end="127">Hitler</Token>
```


Figure 3: XML output example

3. LICENCE

a) The web service only is licenced NaCTeM Web Service Licence Agreement (standard non-commercial use) – see "U-Compare-Tokenisation-Service-Licence.pdf" in the "licences" directory. Please contact us using the details below if you require a commercial licence.

b) The tools used in the workflow on which the web service is based may have their own licences. The NaCTeM Web Service Licence Agreement does NOT apply to these tools.

4. ADMINISTRATIVE INFORMATION

Contact

For further information, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

5. REFERENCES

Ananiadou, S., Thompson, P., Kano, Y., McNaught, J., Attwood, T. K., Day, P. J. R., Keane, J., Jackson, D. and Pettifer, S.. (2011). Towards Interoperability of European Language Resources. *Ariadne,* 67.

Kontonatsios, G., Korkontzelos, I., Kolluru, B., Thompson, P. and Ananiadou, S. (In Press). Deploying and Sharing U-Compare Workflows as Web Services. *Journal of Biomedical Semantics.*