

U-Compare NP Chunking Service

1. BASIC INFORMATION

Service name

U-Compare NP Chunking Service

Overview and purpose of the tool

This is a web service that identifies noun phrase (NP) chunks in Romanian text. Also carries out sentence splitting, tokenisation and POS tagging.

A short description of the algorithm

This web service is based on a UIMA-based workflow, created using the U-Compare text mining system¹. The workflow was exported from U-Compare as a web service using the built in functionality (Kontonastios 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-compliant tools

- 1) MLRS Sentence Splitter (University of Malta)²,
- 2) UAIC-POSTagger (Alexandru Ioan Cuza University, Romania)
- 3) UAIC-NP-Chunker (Alexandru Ioan Cuza University, Romania)

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/NPChunking_UAIC-NPChunker

¹ <http://nactem.ac.uk/ucompare/>

² <http://mlrs.research.um.edu.mt/>

The web form available at this URL is shown in Figure 1, with some Romanian text entered into the text box.

Examples

XML document inline XML stand-off annotation

Prima și cea mai importantă instituție a Republicii Romane era Senatul Roman. În Senat existau două partide neoficiale: optimateș și populares. Senatul a avut o importanță majoră, iar prestigiul său s-a format prin prisma participării instituționale a patricienilor bogați, aparținând aristocrației, și a plebeilor. Romanii respectau două principii pentru oficialii lor: anualitatea sau durata de un an a mandatelor, și colegialitatea sau deținerea aceleiași funcții simultan de către cel puțin două persoane. Statutul suprem de consul, de exemplu, era întotdeauna deținut de două persoane în același timp, fiecare dintre ele exercitând o putere mutuală de veto asupra oricăror acțiuni ale celuilalt consul. Dacă de exemplu întreaga Armată Romană ieșea pe câmpul de luptă, era întotdeauna sub comanda celor doi consuli, care alternau zilele de comandă. Majoritatea celorlalte funcții erau deținute de mai mult de două persoane; în Republica târzie existau 8 pretori în fiecare an și 20 chestori.

Run

Service Description

Web service created by exporting UIMA-based workflow from the U-Compare text mining system.
Functionality: Identifies NP chunks in plain text. Also carries out sentence splitting, tokenization and POS tagging.

Usage

POST request should be sent to use the service

1) text -- the value of this parameter is the text to analyze. Expected encoding is UTF-8

References

Application programming interface

```
String text = "Hello Mr. John Smith !";  
String parameters = "text=" + URLEncoder.encode(text, "UTF-8") +  
"&mode=inline";  
URL url = new URL(url of the service);
```

Figure 1: Web form for the web 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 be 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/NPChunking_UAIC-NPChunker");
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.

The screenshot shows a web interface with a blue header. Below the header, there is a section titled "Select type of annotation" with three radio buttons: "NpChunkWithHead" (checked), "Sentence", and "RichToken". Below this, a text snippet is displayed with various words highlighted in yellow. Below the text, there is a table with the following content:

NpChunkWithHead
cea mai importantă instituție a Republicii Romane
Republicii Romane
Senatul Roman
Roman
Senat
două partide neoficiale
optimates
populares
Senatul
o importanță majoră, iar prestigiul său
prestigiul său
prima participării instituționale
participării instituționale
patricienilor bogați
aristocrației
plebeilor
Romani
două principii pentru oficialii lor
o oficialii lor
anualitatea
durata de un an a mandatelor
un an
mandatelor
colegialitatea

Figure 2: Visualisation of web service output

In Figure 2, the top of the screen has check boxes corresponding to each type of annotation produced by the workflow – in this case “NPChunkWithHead” (corresponding to the NP chunks), “Sentence” and “RichToken” annotations (the latter of which allow both part-of-speech tags and lemma information to be associated with tokens). 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 the “NPChunkWithHead” 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, shows the table of NP chunks, with each row displaying the text covered by the NP chunk. Information about the Sentence and RichToken annotations are displayed further down the page.

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. For the “RichToken” type, the “posString” and “base” attributes of the annotations encode the part-of-speech and lemma of each token, respectively.

```
- <result>
- <Sentence begin="0" end="77">
- <Sentence begin="0" end="77">
  - <NPChunkWithHead begin="9" end="58">
    <RichToken begin="9" end="12" posString="Tdfsr" base="cel">cea</RichToken>
    <RichToken begin="13" end="16" posString="Rg" base="mai">mai</RichToken>
    <RichToken begin="17" end="27" posString="Afpfsm" base="important">importantă</RichToken>
    <RichToken begin="28" end="38" posString="Ncfsm" base="instituție">instituție</RichToken>
    <RichToken begin="39" end="40" posString="Tsfsr" base="a">a</RichToken>
  - <NPChunkWithHead begin="41" end="58">
    <RichToken begin="41" end="51" posString="Ncfsoy" base="republică">Republicii</RichToken>
    <RichToken begin="52" end="58" posString="Afpfson" base="roman">Romane</RichToken>
  </NPChunkWithHead>
</NPChunkWithHead>
<RichToken begin="59" end="62" posString="Vmii3s" base="fi">era</RichToken>
- <NPChunkWithHead begin="63" end="76">
  <RichToken begin="63" end="70" posString="Ncmsry" base="senat">Senatul</RichToken>
  - <NPChunkWithHead begin="71" end="76">
    <RichToken begin="71" end="76" posString="Npmsry" base="roman">Roman</RichToken>
  </NPChunkWithHead>
</NPChunkWithHead>
<RichToken begin="76" end="77" posString="PERIOD" base=".">.</RichToken>
</Sentence>
</Sentence>
```

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-NP-Chunking-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*.