U-Compare Sentence Splitting Service

1. BASIC INFORMATION

Service name

U-Compare Sentence Splitting Service

Overview and purpose of the tool

This is a web service that identifies sentences in Welsh 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 a single UIMA compliant tool:

1) Freeling[2] sentence splitter web service[3], configured to operate on Welsh (service provided by the PANACEA project[4])

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/Sentence_Splitting_Freeling

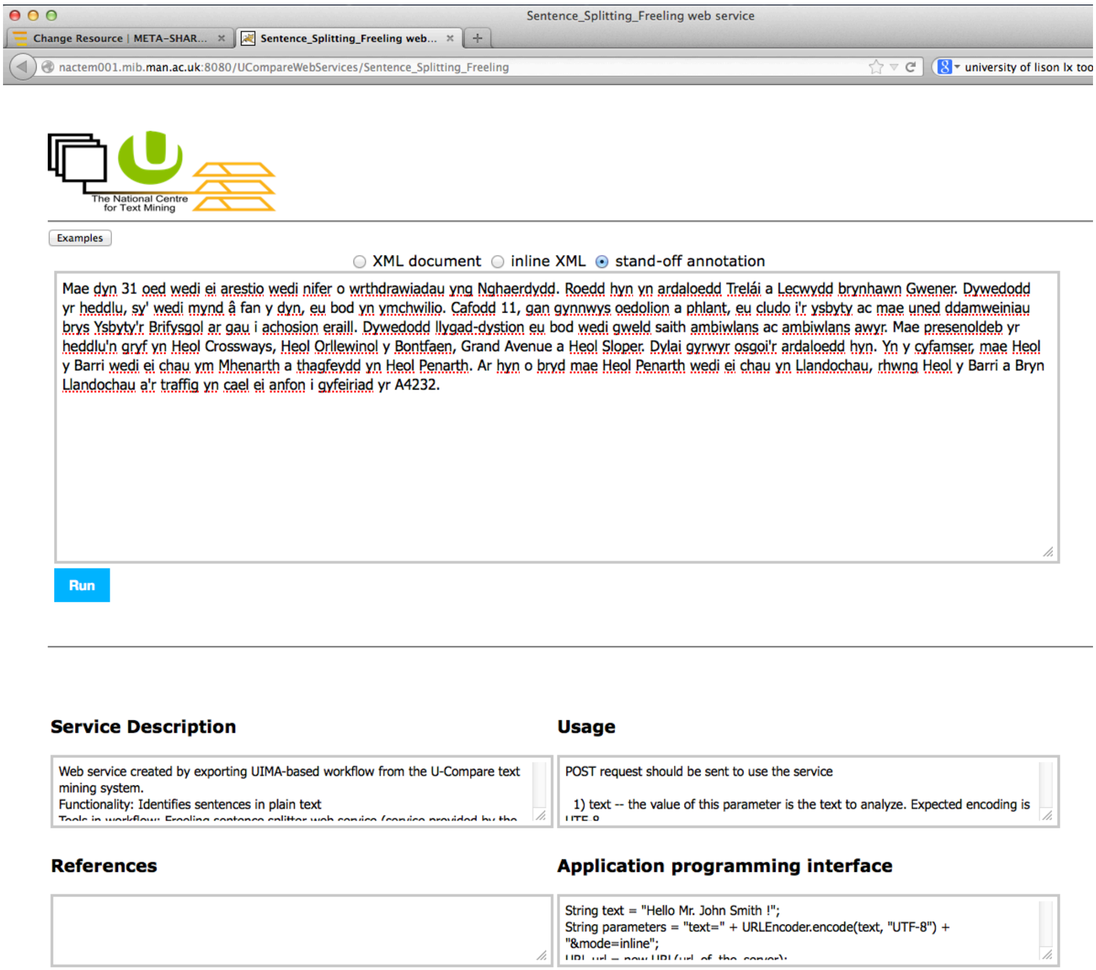
[1] http://nactem.ac.uk/ucompare/

[2] http://nlp.lsi.upc.edu/freeling/

[3] http://registry.elda.org/services/205

[4] http://www.panacea-lr.eu/

The web form available at this URL is shown in Figure 1, with some welsh text entered.


Sentence_Splitting_Freeling web service
Change Resource | META-SHAR...
Sentence_Splitting_Freeling web...
nactem001.mib.man.ac.uk:8080/UCompareWebServices/Sentence_Splitting_Freeling
university of lison lx too
The National Centre for Text Mining
Examples
Run

○ XML document ○ inline XML ● stand-off annotation

Mae dyn 31 oed wedi ei arestio wedi nifer o wrthdrawiadau yng Nghaerdydd. Roedd hyn yn ardaloedd Trelái a Lecwydd brynhawn Gwener. Dywedodd yr heddlu, sy' wedi mynd â fan y dyn, eu bod yn ymchwilio. Cafodd 11, gan gynnwys oedolion a phlant, eu cludo i'r ysbyty ac mae uned ddamweiniau brys Ysbyty'r Brifysgol ar gau i achosion eraill. Dywedodd llygad-dystion eu bod wedi gweld saith ambiwlans ac ambiwlans awyr. Mae presenoldeb yr heddlu'n gryf yn Heol Crossways, Heol Orllewinol y Bontfaen, Grand Avenue a Heol Sloper. Dylai gyrwyr osgoi'r ardaloedd hyn. Yn y cyfamser, mae Heol y Barri wedi ei chau ym Mhenarth a thagfeydd yn Heol Penarth. Ar hyn o bryd mae Heol Penarth wedi ei chau yn Llandochau, rhwng Heol y Barri a Bryn Llandochau a'r traffig yn cael ei anfon i gyfeiriad yr A4232.

Service Description

Web service created by exporting UIMA-based workflow from the U-Compare text mining system.
Functionality: Identifies sentences in plain text
Tools in workflow: Freeling sentence splitter web service (service provided by the

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_server);
```

Figure 1: Web form for the 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/UCompareWebServic
es/Sentence_Splitting_Freeling);
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.

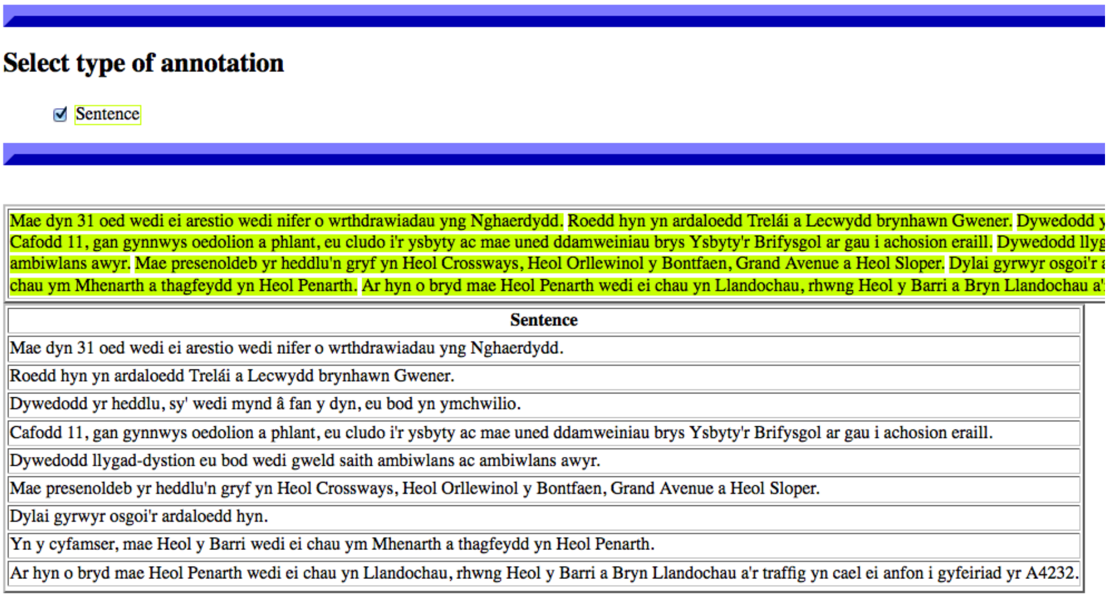
If the service is run programmatically, then the output is provided in XML format. See section 3 for an example.

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

Mae dyn 31 oed wedi ei arestio wedi nifer o wrthdrawiadau yng Nghaerdydd. Roedd hyn yn ardaloedd Treláí a Lecwydd brynhawn Gwener. Dywedodd y
Cafodd 11, gan gynnwys oedolion a phlant, eu cludo i'r ysbyty ac mae uned ddamweiniau brys Ysbyty'r Brifysgol ar gau i achosion eraill. Dywedodd llyg
ambiwlans awyr. Mae presenoldeb yr heddlu'n gryf yn Heol Crossways, Heol Orllewinol y Bontfaen, Grand Avenue a Heol Sloper. Dylai gyrwyr osgoi'r a
chau ym Mhenarth a thagfeydd yn Heol Penarth. Ar hyn o bryd mae Heol Penarth wedi ei chau yn Llandochau, rhwng Heol y Barri a Bryn Llandochau a'

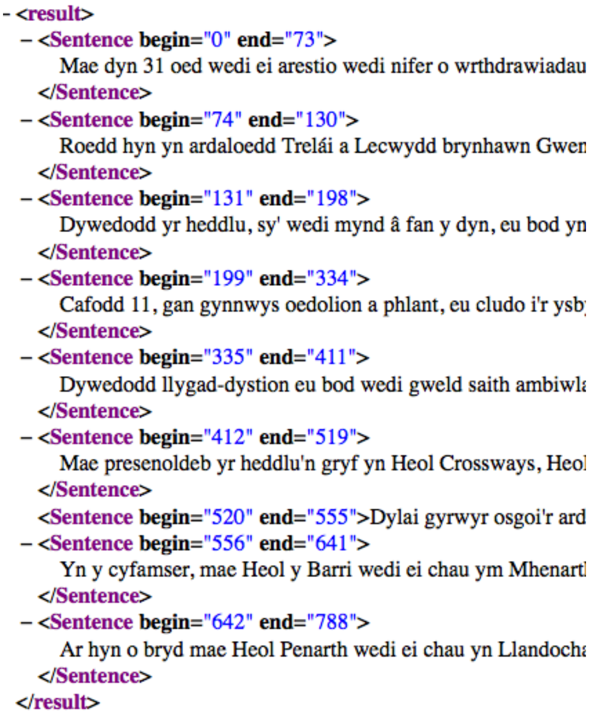
Sentence
Mae dyn 31 oed wedi ei arestio wedi nifer o wrthdrawiadau yng Nghaerdydd.
Roedd hyn yn ardaloedd Treláí a Lecwydd brynhawn Gwener.
Dywedodd yr heddlu, sy' wedi mynd â fan y dyn, eu bod yn ymchwilio.
Cafodd 11, gan gynnwys oedolion a phlant, eu cludo i'r ysbyty ac mae uned ddamweiniau brys Ysbyty'r Brifysgol ar gau i achosion eraill.
Dywedodd llygad-dystion eu bod wedi gweld saith ambiwlans ac ambiwlans awyr.
Mae presenoldeb yr heddlu'n gryf yn Heol Crossways, Heol Orllewinol y Bontfaen, Grand Avenue a Heol Sloper.
Dylai gyrwyr osgoi'r ardaloedd hyn.
Yn y cyfamser, mae Heol y Barri wedi ei chau ym Mhenarth a thagfeydd yn Heol Penarth.
Ar hyn o bryd mae Heol Penarth wedi ei chau yn Llandochau, rhwng Heol y Barri a Bryn Llandochau a'r traffig yn cael ei anfon i gyfeiriad yr A4232.

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 only “Sentence” annotations are present. Checking the box determines whether the sentence annotations are highlighted in the text below.

Below the text, the annotations added by the workflow are shown in tabular format. In Figure 2, there is a single table, which indicates the span of text covered by each "Sentence" annotation (one sentence per 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>
  – <Sentence begin="0" end="73">
      Mae dyn 31 oed wedi ei arestio wedi nifer o wrthdrawiadau
    </Sentence>
  – <Sentence begin="74" end="130">
      Roedd hyn yn ardaloedd Treláí a Lecwydd brynhawn Gwen
    </Sentence>
  – <Sentence begin="131" end="198">
      Dywedodd yr heddlu, sy' wedi mynd â fan y dyn, eu bod yn
    </Sentence>
  – <Sentence begin="199" end="334">
      Cafodd 11, gan gynnwys oedolion a phlant, eu cludo i'r ysb
    </Sentence>
  – <Sentence begin="335" end="411">
      Dywedodd llygad-dystion eu bod wedi gweld saith ambiwl
    </Sentence>
  – <Sentence begin="412" end="519">
      Mae presenoldeb yr heddlu'n gryf yn Heol Crossways, Heol
    </Sentence>
    <Sentence begin="520" end="555">Dylai gyrwyr osgoi'r ard
  – <Sentence begin="556" end="641">
      Yn y cyfamser, mae Heol y Barri wedi ei chau ym Mhenart
    </Sentence>
  – <Sentence begin="642" end="788">
      Ar hyn o bryd mae Heol Penarth wedi ei chau yn Llandoch
    </Sentence>
  </result>
```

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-Sentence-Splitting-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.*