UIMA/U-Compare Apertium Morphological Analyser

1. BASIC INFORMATION

Tool name

UIMA/U-Compare Apertium Morphological Analyser

Overview and purpose of the tool

This tool performs tokenization of text and assigns all possible morphological analyses to each token. These analyses include the base form of the token, part-of-speech, information about number and gender. The morphological analyser is a module of Apertium machine translation system[1] (Armentano-Ollet et al., 2006). The provided tool can currently operate on a subset of the languages that are supported by the Apertium system, namely: English, Spanish, Calatan, Galician, Portuguese, Romanian and Basque.

The tool is provided as a UIMA[2] (Ferrucci et al., 2006) component, specifically as Java archive (jar) file, which can be incorporated within any UIMA workflow. However, it is particularly designed use in the U-Compare text mining plaform (Kano et al., 2009; Kano et al., 2011; see separate META-SHARE record), since the types of annotations it produces are compliant with the U-Compare.

A short description of the algorithm

The morphological analysis is carried using finite-state transducers, in conjunction with morphological dictionaries. See the Apertium documentation (http://wiki.apertium.org/wiki/Documentation) for more information.

2. TECHNICAL INFORMATION

Software dependencies and system requirements

The tool can most easily be run using the U-Compare platform (see separate META-SHARE record). The only requirement to run U-Compare is for Java 6 to be installed.

To run the tool as a UIMA component independently of U-Compare, Apache UIMA must be installed (see http://uima.apache.org/).

[1] http://www.apertium.org/
[2] http://uima.apache.org/

Installation

In order to run the Apertium Morphological analyser in U-Compare, it must be imported into U-Compare. Information about downloading and running U-Compare can be found in the documentation associated with the META-SHARE record for the U-Compare Workbench or the U-Compare website: http://nactem.ac.uk/ucompare/.

Importing the UIMA component (provided as the file ApertiumMorpho.jar) is carried out in U-Compare as follows:

1) From the "Library" menu in the U-Compare Workbench, choose the item "Register External Components (Edit Classpath)". This will cause a window to appear that allows external components to be managed. It is shown in Figure 1.

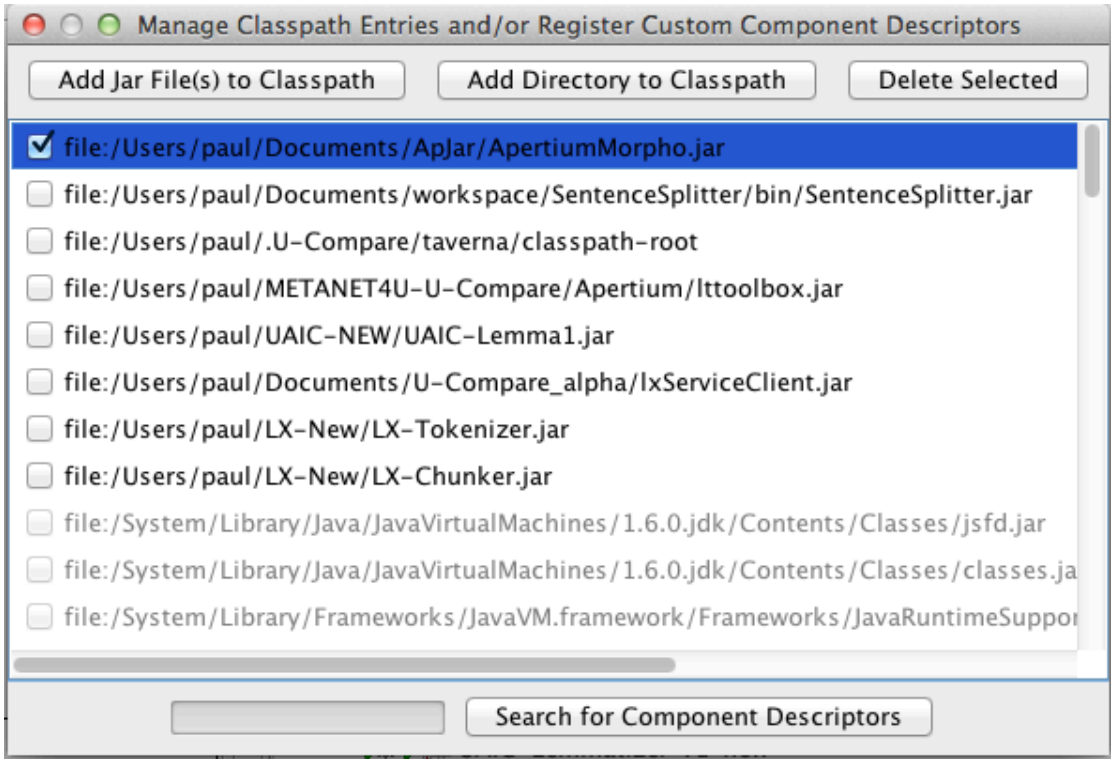

Figure 1: External component management window

2) Click on the button "Add Jar File(s) to Classpath", and browse to the location where the file "ApertiumMorpho.jar" has been saved, and click on "Open". The will cause the file to be displayed in the list of files in the external component manager window.

3) Ensure that box next to the file "ApertiumMorpho.jar" is checked in the external component manager window. Then, click the "Search for Component Descriptors" button. A "Component Descriptors Search Results" window will appear (Figure 2).

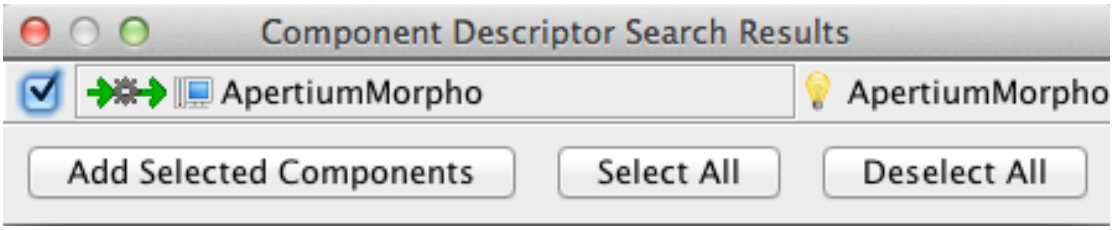

Figure 2: Component Descriptor Search Results window

4) Check the box next to "ApertiumMorpho", and click on the "Add Selected Components" button. The name of the component will then appear in the library on components on the right hand side of the main U-Compare workbench window, under "Custom components", and it can then be used in workflows.

Execution instructions

Within U-Compare, the tool can be executed through inclusion in workflow. This can be done simply by dragging and dropping it onto the workflow canvas using the graphical user interface of the U-Compare workbench. See the META-SHARE record "U-Compare Workbench" for more details. Alternatively, it can be incorporated into other UIMA-based workflows, by following the documentation on the Apache UIMA site: http://uima.apache.org/

The component must be configured before use, to tell the component the languages in which the text(s) to be morphologically analysed are written. This is set by specifying a value for the "languagePair" parameter. Since this tool is a module of a machine translation system, the language data is stored in pairs, i.e., the source language and the target language. The value of the "languagePair" parameter consists of two-letter codes for the languages, joined with a hyphen, where the first language is the source language, and the second is the target language, e.g. "pt-es" is used when Portuguese is the source language and Spanish is the target language.

If this morphological analyser is run in a workflow without the translation module, only the source language is relevant, but the a complete language pair string must still be specified as the value of the "languagePair" parameter, where the language to be analysed appears first in the pair string.

Possible values of the languagePair attribute that can currently be used are as follows: "en-es", "es-en", "gl-es", "es-gl", "es-pt", "pt-es", "es-ca", "ca-es" and "eu-es". If a non-valid value is entered, then the language pair will default to "en-es", i.e,. morphological analysis for English will be carried out

When being run within U-Compare, the value of the "languagePair" parameter can be set by clicking on the  icon. This will cause a parameter configuration window to appear, allowing the user to enter the appropriate language pair string. A part of this

window is shown in Figure 3. The value entered means that the component will be configured to carry out morphological analysis for Spanish.

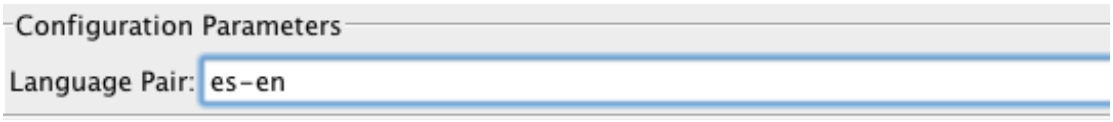

Figure 3: Configuration of LanguagePair parameter

Input/Output data formats

Input data formats

The tool operates on plain, unannotated text. Thus, the UIMA Common Analysis Structure (CAS) should contain the text to be analysed prior to the tool being executed. In a UIMA workflow, this could be achieved by reading in a single text or corpus of text. For example, U-Compare provides collection readers that can read in text from an input box, or otherwise read a directory of texts.

Output data format

An annotation is thus added to the CAS corresponding to each token in a document, with the type "ApertiumToken". This provides a "morphology" field, in which the possible morphological analyses generated for each token are stored. An example of one of these morphological analyses (for Spanish) is as follows:

```
mide/medir<vblex><pri><p3><sg>/medir<vblex><imp><p2><sg>
```

The line begins with the surface form of the token as it appears in the text being analysed. Each possible morphological analysis is separated by a forward slash. The first item on each analysis is the base form, followed by different tags providing morphological information, each enclosed in angled brackets. The first of these is a part-of-speech tag. The other tags will vary according to the part of speech, but in the above example, consist of tense, person and number.

Different CAS consumers (such as those provided in U-Compare) can be used to write the contents of the CAS to a file or database format.

Integration with external tools

As mentioned above, the tool can only be run within U-Compare framework, or more generally, within the UIMA framework.

3. CONTENT INFORMATION

Figure 4 shows the output of the tool in the U-Compare workbench. Each token recognised is separately underlined. The sample text is taken from the CNN Español site (http://www.cnnespanol.com).

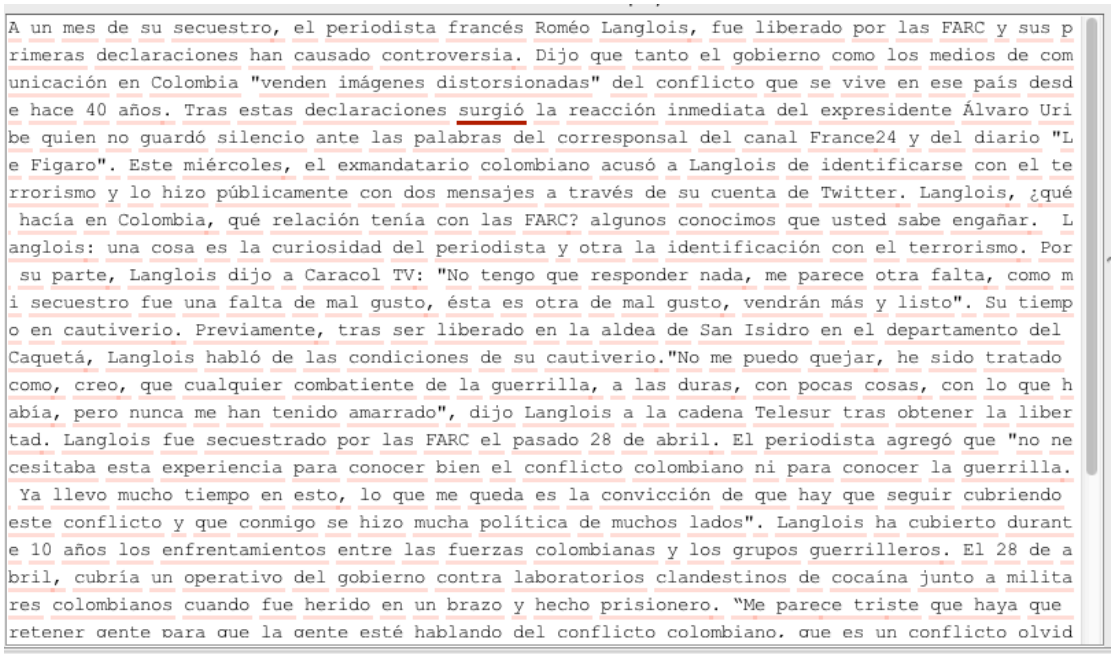

Figure 1: Output of the Apertium morphological analyser in the U-Compare workbench, showing the individual tokens identified

Figure 2 shows another part of analysis displayed in U-Compare, with the attributes of each annotation. These consist of the start and end offsets of each token, together with the different possible morphological analyses for each of them. Words that are unknown by the system are marked with a "*".

begin	end	posString	base	morphology
0	1			A/A<pr>
2	4			un/uno<det><ind><m><sg>
5	8			mes/mes<n><m><sg>
9	11			de/de<pr>
12	14			su/suyo<det><pos><mf><sg>
15	24			secuestro/secuestro<n><m><sg>/secuestrar<vblex><pri><p1><sg>
24	25			,/,<cm>
26	28			el/el<det><def><m><sg>
29	39			periodista/periodista<n><mf><sg>
40	47			francés/francés<n><m><sg>/francés<adj><m><sg>
48	53			Roméo/*Roméo
54	62			Langlois/*Langlois
62	63			,/,<cm>
64	67			fue/ir<vblex><ifi><p3><sg>/ser<vbser><ifi><p3><sg>
68	76			liberado/liberar<vblex><pp><m><sg>
77	80			por/por<pr>
81	84			las/el<det><def><f><pl>/prpers<prn><pro><p3><f><pl>
85	89			FARC/FARC<n><acr><f><pl>
90	91			y/y<cnjcoo>
92	95			sus/suyo<det><pos><mf><pl>
96	104			primeras/primero<adj><f><pl>/primer<det><ord><f><pl>
105	118			declaraciones/declaración<n><f><pl>
119	122			han/haber<vbhaver><pri><p3><pl>
123	130			causado/causar<vblex><pp><m><sg>
131	143			controversia/controversia<n><f><sg>
143	144			./.<sent>
145	149			Dijo/Decir<vblex><ifi><p3><sg>
150	153			que/que<cnjcoo>/que<cnjsub>/que<rel><an><mf><sp>

Figure 2: Attributes for the ApertiumToken annotations, displaying the beginning and end offsets of each token, plus the possible morphological anlyses

Running the tool on the 3 KB Spanish text shown in Figure 1 on a single core machine with 8 GB RAM takes around 0.25 seconds.

4. LICENCES

a) The ApertiumMorpho UIMA wrapper is licensed using the GNU General Public License version 2.0 (GPLv2). Please see "COPYING.txt" in the "Licences" directory. Please acknowledge the National Centre for Text Mining, University of Manchester if you use the ApertiumMorpho UIMA component

b) The underlying Apertium software is licensed using the GNU General Public License version 2.0 (GPLv2). Please see "COPYING.txt" in the "Licences" directory.

c) The UIMA framework is licenced using the Apache licence. Please see "Apache.txt" in the licenses directory.

5. ADMINISTRATIVE INFORMATION

Contact

Contacts for the Apertium system can be found here:
http://wiki.apertium.org/wiki/Contact

For further information regarding this UIMA wrapper for the Apertium morphological analyser module, please contact Sophia Ananiadou:
sophia.ananiadou@manchester.ac.uk

6. REFERENCES

Armentano-Oller, C., Carrasco, R., Corbí-Bellot, A.,Forcada, M., Ginestí-Rosell, M., Ortiz-Rojas, S. (2006). Open-source Portuguese–Spanish machine translation. Computational Processing of the Portuguese Language ,50-59

Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E.W. , Hampp T., Doganata, Y., Welty, C., Amini, L., Kofman, G., Kozakov, L. and Mass, Y. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report RC24122.

Kano, Y., Baumgartner Jr., W. A, McCrochon, L., Ananiadou, S., Cohen, K. B., Hunter, L. and Tsujii, J. (2009). U-Compare: share and compare text mining tools with UIMA. *Bioinfomatics*, 25(15), 1997-1998.

Kano, Y., Miwa, M., Cohen, K. B., Hunter, L., Ananiadou, S. and Tsujii, J.. (2011). U-Compare: a modular NLP workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3), 11:1 - 11:10.