

Labeling Wikinews articles with the Corpus Server and the UIMA Cas Editor

Introduction

This guide will explain on how to do a labeling project with the Corpus Server, the Apache UIMA Cas Editor with the OpenNLP plugins.

Installing

Corpus Server

The Corpus server is running inside an OSGi Runtime Container. In this tutorial Apache Karaf will be used.

Install Apache Karaf

To install Apache Karaf follow the these steps:

- Download it from <http://karaf.apache.org/> (testing here with 2.2.7 and 3.0.0)
- Unpack it
- Start it up with bin/karaf

Build the Corpus Server

Checkout the Corpus Server from svn with these two commands:

- svn co <https://svn.apache.org/repos/asf/opennlp/sandbox/corpus-server>
- svn co <https://svn.apache.org/repos/asf/opennlp/sandbox/corpus-server-impl>

This will create a folder for each maven module.

Go first to the corpus-server, and do a mvn install there,
afterward do the same in the corpus-server-impl module.

Install the Corpus Server

Starting Apache Karaf via bin/karaf open the karaf console.

The following command will install all dependencies and the Corpus Server:

- If using Karaf 2.x
 - features:refreshUrl file:///home/xyz/dev/opennlp/sandbox/corpus-server/feature.xml
 - features:install opennlp-corpus-server
- If using Karaf 3.x
 - feature:repo-refresh file:///home/xyz/dev/opennlp/sandbox/corpus-server/feature.xml
 - feature:install opennlp-corpus-server

The corpus-server itself is just an interface layer which can expose an actual
Corpus Server implementation via its REST API to vairous tools which know this API.

To use the server an actual implementation must be installed in the OSGi runtime as
well, e.g the default implementation corpus-server-impl or a self made one.

This can be done with these commands:

- If using Karaf 2.x
 - features:refreshUrl file:///home/xyz/dev/opennlp/sandbox/corpus-server-impl/feature.xml
 - features:install opennlp-corpus-server-impl
- If using Karaf 3.x
 - feature:repo-refreshfile:///home/xyz/dev/opennlp/sandbox/corpus-server-impl/feature.xml
 - feature:install opennlp-corpus-server-impl

The last command can take several minutes to complete.

The "list" command will now show the Corpus Server:

```
...
[ 76] [Active ] [      ] [ 60] OpenNLP Corpus Server (0.0.1.SNAPSHOT)
[ 77] [Active ] [      ] [ 60] Apache Derby 10.8 (10.8.1000002.1095077)
[ 78] [Active ] [      ] [ 60] OpenNLP Corpus Server Implementation (0.0.1.SNAPSHOT)
```

Apache UIMA Cas Editor

- Download and install the latest eclipse 3.x version
- Install the latest Apache UIMA Cas Editor (See "UIMA Overview & SDK Setup" in the UIMA documentation) (<http://uima.apache.org/documentation.html>)

Loading the Wikinews Corpus

Get the Wikinews dumps

The wikinews dumps can be downloaded from Wikipedia, here is some general information about the dumps:
<http://meta.wikimedia.org/wiki/Data.dumps>

The dump itself can be downloaded here: <http://dumps.wikimedia.org/>

Choose a mirror near you and then go to "Database backup dumps". The file you need to download is called like this (the date will be different): enwikinews-20120727-pages-articles.xml.bz2

After the download decompress the file, e.g. with bunzip2 on Linux.

Convert the dump files to Apache UIMA XMI files

The current version of the parser only works well for the English wikinews dump. Contributions to fix this for other languages are very welcome.

Get and compile the Wikinews Importer

Checkout the wikinews parser:

svn co <https://svn.apache.org/repos/asf/opennlp/sandbox/wikinews-importer/>

And compile it with this command: mvn clean install

Parse the XML articles

The xml file can now be parsed:

bin/convertor /home/blue/Downloads/enwikinews-20120727-pages-articles.xml articles

This command will take a while to run, when its done there is one xmi file for each article in the articles folder.

Load the articles to the Corpus Server

To load the articles in the corpus server a corpus must be created first.
This is done with the corpus-server-tools.

Get and compile the Corpus Server Tools

Checkout the tools

svn co <https://svn.apache.org/repos/asf/opennlp/sandbox/corpus-server-tools>
and build them with mvn clean install.

Create a new Corpus

Now create the wikinews corpus in the previously started Corpus Server
can be created:

bin/cs-tools CreateCorpus <http://localhost:8080/rest/corpora> enwikinews .../wikinews-importer/samples/TypeSystem.xml .../wikinews-importer/samples/wikinews.xml

The response code should be 204. If something goes wrong you get an HTTP error code.

Import the articles to the created corpus

And import the article files:

bin/cs-tools CASImporter <http://localhost:8080/rest/corpora/enwikinews> .../wikinews-importer/articles

Opening an article in the Cas Editor

Setting up the OpenNLP annotator plugin

Training a component on the data