Documentation for

JULIELAB UIMA Wrapper for OpenNLP Part-of-Speech (POS) Tagger

Ekaterina Buyko
Jena University Language & Information Engineering (JULIE) Lab
Fürstengraben 30
D-07743 Jena, Germany
{buyko}@coling-uni-jena.de

1 Objective

The OpenNLP POS Tagger[1] provides part of speech tags for tokens. JULIELAB UIMA Wrapper for OpenNLP POS Tagger is part of the JULIE NLP tool suite[2] which contains several NLP components (all UIMA compliant) from sentence splitting to named entity recognition and normalization as well as a comprehensive UIMA type system.

2 Requirements and Dependencies

JULIELAB UIMA Wrapper for OpenNLP POS Tagger is written in Java 1.5 using Apache UIMA version 2.1.0-incubation[3]. It was not tested with other UIMA versions.

The input and output of an AE takes place by annotation objects. The classes corresponding to these objects are part of a *JULIE UIMA Type System*[4].

[1] http://www.opennlp.org

[2] http://www.julielab.de/

[3] http://incubator.apache.org/uima/

[4] The *JULIE UIMA type system* can be obtained from http://www.julielab.de/

3 Using the AE – Descriptor Configuration

In UIMA, each component is configured by a descriptor in XML. In the following we describe how the descriptor required by this AE can be created with *Component Descriptor Editor*, an Eclipse plugin which is part of the UIMA SDK.

A descriptor contains information on different aspects. The following subsection refers to each sub aspect of the descriptor which is, in the Component Descriptor Editor, a separate *tabbed page*. For an indepth description of the respective configuration aspects or tabs, please refer to the *UIMA SKD User's Guide*[5], especially chapter 12 on “Component Descriptor Editor User's Guide”.

To define your descriptor go through each tabbed pages mentioned here, make your respective entries (especially in page *Parameter Settings* you will be able to configure OpenNLP POS Tagger to your needs) and save the descriptor as `PosTagAnnotator.xml`.

Overview This tab provides general informtion about the component. For the OpenNLP Sentence Splitter you need to provide the information as specified in Table 1.

Subsection	Key	Value
Implementation Details	Implementation Language	Java
	Engine Type	Primitive
Runtime Information	updates the CAS	yes
	multiple deployment allowed	yes
	outputs new CASes	no
	Name of the Java class file	`de.julielab.jules.ae.opennlp.PosTagAnnotator`
Overall Identification Information	Name	JULIELAB UIMA Wrapper for OpenNLP Tokenizer
	Version	2.0
	Vendor	julielab
	Description	see above

Table 1: Overview/General Settings for AE.

Aggregate Not needed here, as this AE is a primitive.

[5] `http://incubator.apache.org/uima/`

Parameters See Table 2 for a specification of the configuration parameters of this AE. Do not check "Use Parameter Groups" in this tab.

Parameter Name	Parameter Type	Mandatory	Multivalued	Description
modelFile	String	yes	no	path to the OpenNLP POS Tagger model.
tagset	String	no	no	CAS types to annotate (see *JULIE UIMA type system*)
language	String	yes	no	language (e.g. eng) (see *JULIE UIMA type system*)
useTagdict	Boolean	yes	no	`true` if a tag dictionary should be used
tagDict	String	yes	no	path to a tag dictionary (if the parameter value of useTagdict is `true`)
caseSensitive	Boolean	no	no	`true` if a tag dictionary is case senstive

Table 2: Parameters of this AE.

Parameter Settings The specific parameter settings are filled in here. For each of the parameters defined in 3, add the respective values here (has to be done at least for each parameter that is defined as mandatory). See Table 3 for the respective parameter settings of this AE.

Parameter Name	Parameter Syntax	Example
modelFile	model.bin.gz	resources/POSTaggerPennBio.bin.gz
tagset	CASType	de.julielab.jules.types.PennBioIEPOSTag
language	ISO 639-1/2	en
useTagdict	Boolean	yes
tagDict	dir/tagdict	resources/tagdictPennBioIE
caseSensitive	Boolean	no

Table 3: Parameter settings of this AE.

Type System On this page, go to *Imported Type* and add the *julie-morpho-syntax-types.xml* type system. (Use "Import by Location").

Capabilities Nothing needs to be done here.

Index Nothing needs to be done here.

Resources Nothing needs to be done here.

4 Copyright and License