Session 2
Marking-up Provenance in an eScience environment for Linguists
Thorsten Trippel, Claus Zinn

RDA-Deutschland-Treffen 2015
Potsdam, 25.-26.11.2015

research data sharing without barriers
rd-alliance.org
Linguistics and eScience

- linguistics: a theoretical, descriptive, applied and experimental discipline
- community involved in eScience infrastructures
  - CLARIN’s Virtual Language Observatory offers information about ca. 870,000 language resources
  - CLARIN’s Federated Content Search offers search across corpora collections
- many resources can be accessed online
- large number of tools available to process resources
- Weblicht as eScience environment
environment for automatic annotation of text corpora
  - tokenization, part-of-speech tagging, syntactic analysis, named-entity recognition
  - for German, English, and other languages
tools encapsulated as web services
tool pipelining support
  - easy processing chains
  - customised processing chains
researchers can quickly plug&play
  - No installation efforts
  - hence, limited control over tools
PARIS — The suspected ringleader of the Paris attacks was killed Wednesday in a massive pre-dawn raid by French police commandos, two senior European officials said, after investigators followed leads that the fugitive Islamic State militant was holed up north of the French capital and could be plotting another wave of violence.

More than 100 police officers and
Weblicht

**Mode Selection**

- **Easy Mode**
  - Use a pre-defined chain.
  - Valid Input Types [more...]

- **Advanced Mode**
  - Build a customized chain.
  - Valid Input Types [more...]
Weblicht

Available Annotations for: English Plain Text

- Pos Tags/Lemmas
- Morphology
- Constituent Parses
- Dependency Parses
- Named Entities

Visualization Area

Your results will be visualized here once you have:

1. Chosen one of the annotations on the left
2. Run the tools by clicking Run Tools below

Tip: hovering over an annotation type will give a detailed overview of the output.

Input and Chain Selection

PARIS — The sus [Plain T]
PARIS — The suspected ringleader of the Paris attacks was killed Wednesday in a massive pre-dawn raid by French police commandos, two

SfS: To TCF Converter
Language: English
Document Type: TCF
TCF Version: 0.4
Text

SfS: Stanford Tokenizer
Sentences
Tokens

SfS: Illinois Named Entity
Named Entities: CoNLL-2002

Run Tools  Clear Results  Download chain
The dead included the suspected overseer of the Paris bloodshed, Abdelfatah Abaaoud, according to the two senior European officials. Abaaoud, a Belgian extremist, had once boasted that he could slip easily between the lines.
<cmd:WebServiceToolChain>
  <cmd:GeneralInfo>
    <cmd:Descriptions>
      <cmd:Description/>
    </cmd:Descriptions>
    <cmd:ResourceName>myChain</cmd:ResourceName>
    <cmd:ResourceClass>Toolchain</cmd:ResourceClass>
  </cmd:GeneralInfo>
  <cmd:Toolchain>
    <cmd:ToolInChain>
      <cmd:PID>http://hdl.handle.net/11858/00-1778-0000-0004-BA56-7</cmd:PID>
    </cmd:ToolInChain>
    <cmd:ToolInChain>
      <cmd:PID>http://hdl.handle.net/11022/0000-0000-2518-C</cmd:PID>
    </cmd:ToolInChain>
    <cmd:ToolInChain>
      <cmd:PID>http://hdl.handle.net/11022/0000-0000-839F-9</cmd:PID>
    </cmd:ToolInChain>
  </cmd:Toolchain>
</cmd:WebServiceToolChain>
Weblicht

Illinois Named Entity Recognizer

This is a state of the art NER tagger that tags plain text with named entities. The newest version tags entities with the "classic" 4-label type set (people / organizations / locations / miscellaneous)

Illinois NER labels names in text

HttpContext <webService service="webService" type="TypeOfWebService" RESTfull="RESTfull"

http://weblicht.sfs.uni-tuebingen.de/rws/service-lbjner/annotate/stream

LifecycleStatus development

PublicationDate 2014-07-24T11:57:28.489+02:00

LastUpdate 2015-09-17T22:46:16.561+02:00

Contact wlsupport@sfs.uni-tuebingen.de

Creation

Creators

Organisation SFS: Uni-Tuebingen

Operations

Operation

Default

Input

ParameterGroup

Input Parameters

Parameters

Parameter

type

WebServiceArgValue

AllowManualSelectionFallback false

Values

ParameterValue

text/tcf+xml

Values

ParameterValue

Version
This is a state of the art NER tagger that tags plain text with named entities. The newest version tags entities with the "classic" 4-label type set (people / organizations / locations / miscellaneous).

Creator: SfS: Uni-Tuebingen

Contact: wlsupport@sfs.uni-tuebingen.de

PID: http://hdl.handle.net/11022/0000-0000-839F-9

<table>
<thead>
<tr>
<th>Input Features</th>
<th>Output Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>lang: en</td>
<td>namedentities.type:</td>
</tr>
<tr>
<td>type: text/tcf+xml</td>
<td>version: 0.4</td>
</tr>
<tr>
<td>version: 0.4</td>
<td>conll2002</td>
</tr>
<tr>
<td>sentences</td>
<td></td>
</tr>
<tr>
<td>tokens</td>
<td></td>
</tr>
</tbody>
</table>
Marking up Provenance Information (CLARIN-PLUS)

- switchboard service
  - mediating between resources and tools
  - researchers identifies resource in Virtual Language Observatory or Federated Content Search
  - uses the switchboard to identify all tools to support the analysis of the resource
  - switchboard has all relevant metadata about tools to check whether they are applicable, and if, what types of analyses they provide
  - switchboard sends user to tool
  - but need to guarantee that derived data is decorated with provenance information, but how?
Marking up Provenance Information (CLARIN-PLUS)

- use metadata provenance standards
  - PROV-DM: The PROV Data Model
  - W3C Recommendation 2013
  - has RDF mapping