Uplug

1. BASIC INFORMATION

1.1 Tool name

Uplug.

1.2 Overview and purpose of the tool

Uplug (see Tiedemann, 2003a) is a collection of tools and scripts for processing text-corpora, for automatic alignment and for term extraction from parallel corpora.

Several tools have been integrated in Uplug. Pre-processing tools include a sentence splitter, a general tokenizer and wrappers around external part-of-speech tagger and shallow parsers. The following external tools are included in the standard package: The Grok system for English (tagging and chunking), and the morphological analyzer ChaSen for Japanese. Translated documents can be sentence aligned using the length-based approach by Gale&Church, hunalign or GMA by Melamed and others. Words and phrases can be aligned using the clue alignment approach (see Tiedemann, 2003b) and GIZA++ (a toolbox for training statistical alignment models for SMT). Other tools can easily be integrated, for example, the TreeTagger for English, French, Italian, and German, the TnT tagger for English, German and Swedish.

Uplug has been developed within the PLUG project (see Tiedemann, 2002). It also includes web-based interfaces for interactive sentence and word alignment (see Tiedemann, 2006).

For more information about Uplug, visit `http://stp.lingfil.uu.se/~joerg/Uplug/`.

1.3 A short description of the algorithm

Not applicable.

2. TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

Linux and a recent version of Perl (version 5.8.0).

2.2 Installation

I. Uplug runs out of the box on standard GNU/Linux machines without any particular installation routines. Some external tools need to be compiled for non-standard operating systems. Some scripts need to be adjusted in case Perl is installed in non-standard locations (expected location: /usr/bin/perl). In particular, you may need to look at the following parts:

1. Check you installation of Perl.

You have to modify scripts in bin/ to point to your perl binary

(the default setting is /usr/bin/perl)

2. Some scripts in tools/ use bash (/bin/bash)

You might have to adjust scripts in tools/ according to your installation

3. Check the external programs in ext/

External programs are pre-compiled for 32-bit GNU/Linux (RedHat 9)

Starter scripts have to be modified if necessary:

ext/tagger

ext/parser

ext/chunker

II. Some of the external programs are not included in the package. You might want to install them on your local machine:

The TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

The Qtag Tagger: http://web.bham.ac.uk/O.Mason/software/tagger/

The TnT tagger: http://www.coli.uni-sb.de/~thorsten/tnt/

- The TreeTagger is expected to be in ext/tree-tagger.
- Qtag is expected in ext/Qtag with sub-directories for different languages (check startup scripts in ext/tagger/qtag_XXXXX).
- TnT is expected in ext/tnt with a subdirectory 'models' for the language models (check scripts again).

III. The implementation of the web-applications for interactive alignment can be found in uplug/web/php. Please look at the README file in that directory for information about installation and usage of the tools. The tools are implemented in PHP and require Uplug and a running web server with PHP-support installed. There are two main scripts isa.php (for interactice sentence alignment) and ica.php (for interactive word (clue) alignment). A corpus needs to be prepared on the server-side before running the web-interface. There is a Makefile that can be used to simplify this task. Note that the web-server requires write-permissions for the data directories to store the alignment information.

2.3 Execution instructions

For all detailed informations and instructions about the Uplug, see <Quickstart> at: http://stp.lingfil.uu.se/~joerg/Uplug/. Some additional information is available in the root directory of the Uplug package.

For the interactive alignment tools: Look at uplug/web/php/README within the uplug package.

2.4 Input/Output data formats

Uplug is a modular tool and may process various kinds of data files. A common input format is plain text (in various character encodings) and the ouput is in most cases some kind of XML. Some tools require two input files (for alignment, for example) and some others run a pipe-line of tools on a single document. For more information, look at the documentation within the package.

2.5 Integration with external tools

Not applicable.

3. CONTENT INFORMATION

3.1 A test input file (small.txt)

```
Sweden's policy of neutrality is of decisive importance for our peace and
independence. It also contributes to stability and détente in our part of the
world. There is wide popular support for this policy. It will be pursued with
firmness and consistency.
```

3.2 The output file

Calling Uplug with pre-processing and annotation tools for English with the following command

```
/path/to/uplug pre/en-all -in small.txt -ci 'iso-8859-1' > small.xml
```

produces the following output (basic XML markup, sentence splitting, tokenization, POS tagging with the TreeTagger, Grok and chunking with Grok, assuming that the TreeTagger is installed):

```
<?xml version="1.0" encoding="utf-8"?>
<text>
<p id="1">
<s id="s1.1">
 <chunk type="NP" id="c1.1-1">
  <w tree="NP" lem="Sweden" pos="NNP" id="w1.1.1">Sweden</w>
 </chunk>
 <chunk type="NP" id="c1.1-2">
  <w tree="POS" lem="'s" pos="POS" id="w1.1.2">'s</w>
  <w tree="NN" lem="policy" pos="NN" id="w1.1.3">policy</w>
 </chunk>
 <chunk type="PP" id="c1.1-3">
  <w tree="IN" lem="of" pos="IN" id="w1.1.4">of</w>
 </chunk>
 <chunk type="NP" id="c1.1-4">
  <w tree="NN" lem="neutrality" pos="NN" id="w1.1.5">neutrality</w>
 </chunk>
 <chunk type="VP" id="c1.1-5">
  <w tree="VBZ" lem="be" pos="VBZ" id="w1.1.6">is</w>
 </chunk>
 <chunk type="PP" id="c1.1-6">
  <w tree="IN" lem="of" pos="IN" id="w1.1.7">of</w>
 </chunk>
...
```

Processing pipelines can be created using a simple (Perlish) syntax as a combination of modules available in Uplug. For example, the definition of the pre-processing for English used above looks like

this:

```
{
  'module' => {
    'name' => 'English pre-processing',
    'submodules' => [
        '$UplugSystem/pre/markup',
        '$UplugSystem/pre/sent',
        '$UplugSystem/pre/en/toktag',
        '$UplugSystem/pre/en/tagGrok',
        '$UplugSystem/pre/en/chunk',
    ],
    'submodule names' => [
        'basic XML markup',
        'sentence splitter',
        'English tokenizer+tagger',
        'English tagger (Grok)',
        'English chunker (Grok)',
    ],
    'stdin' => 'text',
    'stdout' => 'text',
  },
'input' => {
    'text' => {
      'format' => 'text',
    }
  },
  'output' => {
    'text' => {
      'format' => 'xml',
      'root' => 's',
      'write_mode' => 'overwrite',
      'status' => 'chunk'
    }
  },
  'arguments' => {
    'shortcuts' => {
       'in' => 'input:text:file',
        'out' => 'output:text:file',
       'ci' => 'input:text:encoding',
       'co' => 'output:text:encoding',
    }
  }
}
```

Standard modules are stored in `/path/to/uplug/systems` and can be called without the absolute path (see the uplug command line call above in which the English pre-processing module `/path/to/uplug/systems/pre/en-all` is called like `pre/en-all`). You may define your own configurations using the same syntax. Call them using the absolute path.

Sentence alignment is stored as stand-off annotation:

```
<linkGrp targType="s" toDoc="fr/1988.xml.gz" fromDoc="en/1988.xml.gz">
...
<link certainty="166" xtargets="s3.1;s3.1" id="SL2.1" />
```

```
<link certainty="17" xtargets="s3.2;s3.2" id="SL2.2" />
<link certainty="217" xtargets="s3.3;s3.3" id="SL2.3" />
<link certainty="33" xtargets="s3.4;s3.4" id="SL2.4" />
```

The comma-separated ID's in the xtarget attribute point to sentences in the linked documents

Word alignment uses the same format but adds lexical links:

```
<link certainty="17" xtargets="s3.2;s3.2" id="SL2.2">
 <wordLink certainty="0.0146391749904005" lexPair="our world;unserem Welt"
xtargets="w3.2.9+w3.2.13;w3.2.9+w3.2.12" />
 <wordLink certainty="0.04" lexPair="stability;Stabilität" xtargets="w3.2.5;w3.2.5" />
 <wordLink certainty="0.161677777777778" lexPair="It;Sie" xtargets="w3.2.1;w3.2.1" />
 <wordLink certainty="0.02" lexPair="part;trägt" xtargets="w3.2.10;w3.2.2" />
 <wordLink certainty="0.0199438120050112" lexPair="to in of the .;zu in Teil der ."
xtargets="w3.2.4+w3.2.8+w3.2.11+w3.2.12+w3.2.14;w3.2.4+w3.2.8+w3.2.10+w3.2.11+w3.2.14" />
 <wordLink certainty="0.109435675813194" lexPair="and;und" xtargets="w3.2.6;w3.2.6" />
</link>
```

The clue alignment approach may use various resources and statistics to optimize the alignment between words. It can run iteratively and may use external tools such as GIZA++. Weights of resources can be adjusted manually.

3.3 Approximation of the time necessary to process the test input file.

Not applicable.

4. ADMINISTRATIVE INFORMATION

4.1 Contact person

Name: Jörg Tiedemann
Address:
Department of Linguistics and Philology
Uppsala University
Box 635
SE-75126 Uppsala/Sweden
Affiliation: Department of Linguistics and Philology – Uppsala University
Position: Visiting Professor
Telephone: +46 (0)18-471 1412
Fax: +46 (0)18-471 1094
E-mail: jorg.tiedemann@lingfil.uu.se

5. LICENSE

6. RELEVANT REFERENCES AND OTHER INFORMATION

Tiedemann, J. (2002). “Uplug - a modular corpus tool for parallel corpora”. In Lars Borin, editor, *Parallel Corpora, Parallel Worlds* - Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22-23 April, 1999. Rodopi, Amsterdam, New York, pp. 181-197.

Tiedemann, J. (2003a). *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*, Doctoral Thesis, Studia Linguistica Upsaliensia 1.

Tiedemann, J. (2003b). Combining Clues for Word Alignment. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL03)* Budapest, Hungary, April, pp. 12-17, 2003.

Tiedemann, J. (2006) ISA & ICA - Two Web Interfaces for Interactive Alignment of Bitexts. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, (LREC'2006)*, 2006.