UplugWeb - Corpus Tools on the Web

Jörg Tiedemann
Department of Linguistics
Uppsala University
Box 527
SE-75120 Uppsala
Sweden
Tel +46 (0)18-471 7007
Fax +46 (0)18-471 1416
joerg@stp.ling.uu.se

15th May 2001

1 Introduction

In this report UplugWeb, a collection of web-based corpus tools, is described. The system applies modules of the Uplug corpus tools that have been developed with a focus on word alignment of parallel texts. UplugWeb provides tools for monolingual texts and for parallel bilingual texts of limited size.

The Uplug system was developed within the co-operative project on parallel corpora PLUG [SH99]. It is a modular system designed for processing text corpora with a focus on parallel texts and word alignment [Tie99a]. Several modules that carry out various tasks of processing textual data have been developed and integrated in the Uplug environment. The main application is the Uppsala Word Aligner, UWA [Tie99b].

The UplugWeb interfaces give external users access to the Uplug corpus tools without a local installation of the system. It allows external users to run UWA processes on texts of limited size and the results of the process are returned to the user on the web, as well as by e-mail.

2 The System

The UplugWeb tools provide interfaces for running sub-systems of UWA, which might be of interest even as stand-alone tasks, such as phrase generation and morphological pattern recognition. UplugWeb also includes interfaces for running the complete word alignment process on small-size bitexts. A module for sentence alignment has been integrated in the Uplug environment as well. The results are sent to the user via e-mail on demand. Time-consuming tasks (such as word alignment) run as background processes. In such cases, the UplugWeb interface prints a message about the processing queue in which the tasks were scheduled and the results are sent by e-mail.

The following tools are included in UplugWeb:

Bilingual Sentence Alignment: The Uplug sentence aligner applies the approach proposed by Gale and Church ([GC93]) modified and adjusted to the requirements at Uppsala University [TKS96]. The sentence aligner takes two plain text files as its input and converts them to sentence aligned bitexts with SGML markup in conformity with the TEI standard [SMB94].

Bilingual Word Alignment: The UplugWeb word aligner applies the Uppsala word aligner (UWA). UWA produces two kinds of output: token links and type links. Token links include all word and phrase instances that have been aligned in the text, i.e. token pairs including information about the origin in the text in form of byte spans and segment identifiers. Type links are compiled from token links and comprise unique word or phrase pairs including their linking frequency.

Generation of Word Collocations: Phrase generation is one of the pre-processing tasks in word alignment by UWA (assuming static segmentation) [Tie00]. However, this task might also be interesting as a stand-alone process. The phrase generation module takes a monolingual plain text file as its input and generates significant word collocations. The process is iterative starting with bigrams. Significance is measured by means of co-occurrence metrics such as Mutual Information, Dice, or t-scores. The generation applies additionally a list of phrase boundary words in order to improve the result. The system will use automatically generated lists of stop word if necessary.

Generation of Morphological Pattern: The generation of morphological patterns is based on investigations of string similarity between wordforms in a monolingual corpus. Based on matching and non-matching parts, the system tries to compute common patterns for various word groups. Similarity measures and types of dissimilarity (suffix, prefix, and/or infix) can be adjusted according to the users needs. Similar words are clustered into (morphological) categories with the affix pattern attached.

Handling Web Processes: The system takes care of the processing queue in such a way that simultaneous tasks do not block each other and that the local web server is not overloaded by the UplugWeb system.

Handling Bilingual Corpora: UplugWeb includes tools for exploring bilingual texts and word alignment results. Submitted bitexts will be stored in the UplugWeb database and the user can search for bilingual concordances in the complete collection. Furthermore, it is possible to "hide" corpora from other users ("private" corpora). Word alignment results will be stored in a general link database that can be searched by the users. UplugWeb includes an interface for exploring word alignments related to their origin.

3 Results

Results of the tools as described above are not presented in this abstract. Texts of different types in several languages have been processed by the UplugWeb toolbox for small-scale tests and evaluation.

4 Conclusions

The UplugWeb interfaces are intended for demonstration purposes. They represent a small-scale version of the Uplug corpus tools. The web-processes are restricted to small-sized corpora and the performance depends on the load on the local web-server. Access to the web-interfaces is restricted by web-authorization as a means of controlling the number of users. The UplugWeb toolbox can be used for educational purposes and for small-size applications.

References

[GC93] W. A. Gale and K. W. Church. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19:75–102, 1993.

[SH99] Anna Sågvall Hein. The PLUG Project: Parallel corpora in Linköping, Uppsala, and Göteborg: Aims and achievements. Technical Report 16, Department of Linguistics, University of Uppsala, 1999.

[SMB94] C.M. Sperberg-McQueen and Lou Burnard. Guidelines for Electronic text Encoding and Interchange. Availabe from http://etext.virginia.edu/TEI.html, 1994.

[Tie99a] Jörg Tiedemann. Automatic Construction of Weighted String Similarity Measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, Department of Linguistics, Uppsala University, Sweden, 1999.

[Tie99b] Jörg Tiedemann. Uplug - a modular corpus tool for parallel corpora. Technical Report 17, Department of Linguistics, University of Uppsala, 1999.

[Tie00] Jörg Tiedemann. Exatracting Phrasal Terms using Bitexts. In *Proceedings of the Workshop on Terminology Resources and Computation, in connection with LREC-2000*, pages 57–63, Athens, Greece, 2000. European Language Resources Association.

[TKS96] Erik F. Tjong Kim Sang. Aligning the Scania Corpus. Technical report, Department of Linguistics, University of Uppsala, 1996.

A URL's

Uplug Main Page

http://stp.ling.uu.se/~joerg/uplug

Links to Private Pages

http://stp.ling.uu.se/~joerg/uplug/PrivateHomeName.html

Sentence Alignment

http://stp.ling.uu.se/~joerg/uplug/SentAlign.html

Word Alignment

http://stp.ling.uu.se/~joerg/uplug/UwaWeb.html
http://stp.ling.uu.se/~joerg/uplug/UwaWeb_tei.html
http://stp.ling.uu.se/~joerg/uplug/UwaWebPro.html
http://stp.ling.uu.se/~joerg/uplug/UwaWebPro_tei.html

Phrase Generation

http://stp.ling.uu.se/~joerg/uplug/GenPhrases.html
http://stp.ling.uu.se/~joerg/uplug/GenPhrasesPro.html

Morphological Pattern Generation

http://stp.ling.uu.se/~joerg/uplug/GenMorph.html
http://stp.ling.uu.se/~joerg/uplug/GenMorphPro.html

Query Tools

http://stp.ling.uu.se/~joerg/uplug/SearchLinks.html
http://stp.ling.uu.se/cgi-bin/joerg/uplug/SearchLinkDB.pl
http://stp.ling.uu.se/cgi-bin/joerg/uplug/UserCorpora.pl