LAT Bridge: Bridging tools for annotation and exploration of rich linguistic data

Marc Kemps-Snijders, Thomas Koller, Han Sloetjes, Huib Verwey
Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
E-mail: Marc.Kemps-Snijders@mpi.nl, Thomas.Koller@mpi.nl, Han.Sloetjes@mpi.nl, Huib.Verwey@mpi.nl

Abstract

We present a software module, the LAT Bridge, which enables bidirectional communication between the annotation and exploration tools developed at the Max Planck Institute for Psycholinguistics as part of our Language Archiving Technology (LAT) tool suite. These existing annotation and exploration tools enable the annotation, enrichment, exploration and archive management of linguistic resources. The user community has expressed the desire to use different combinations of LAT tools in conjunction with each other. The LAT Bridge is designed to cater for a number of basic data interaction scenarios between the LAT annotation and exploration tools. These interaction scenarios (e.g. bootstrapping a wordlist, searching for annotation examples or lexical entries) have been identified in collaboration with researchers at our institute.
We had to take into account that the LAT tools for annotation and exploration represent a heterogeneous application scenario with desktop-installed and web-based tools. Additionally, the LAT Bridge has to work in situations where the Internet is not available or only in an unreliable manner (i.e. with a slow connection or with frequent interruptions). As a result, the LAT Bridge's architecture supports both online and offline communication between the LAT annotation and exploration tools.

1. Introduction

The Max Planck Institute for Psycholinguistics (MPI) has developed a Language Archiving Technology (LAT) tool suite to support annotation, enrichment, exploration and archive management of linguistic resources. The supported resource types include annotated media files, lexica, audio, video and image resources. The LAT tools are both used internally by the researchers at the MPI as well as by many others worldwide.

Bootstrap a wordlist:
A user has made a set of annotations and wants to bootstrap a wordlist from this set of annotation resources which acts as the nucleus for a new lexicon. The information from selected tiers is gathered and inserted into the new lexicon.

Search for an Annotation Example:
A user is using a lexicon and wants to look up an example. She wants to search through a set of annotations to locate the appropriate fragment and wants to be able to start that fragment from within the lexical entry using ANNEX. The selection process should be able to transparently supply all necessary information such as time span information to display the right fragment.

Search for a Lexical Entry:
A user is working on an annotation and wants to lookup the corresponding lexical entry which LEXUS should show. The user opens an annotation file in her annotation tool. The user may select a fragment from a tier to search for in the lexicon tool. Alternatively, the user may supply the search term for the lexicon tool to conduct the search on. The information returned from the lexicon may be added as information to the annotation to supplement already existing annotation information.

Word Completion and Correction:
A user is creating an annotation and wants to use lexical knowledge in the form of completion or correction. While the user enters the annotation information, the lexicon is automatically checked to determine whether the information is already available in the lexicon. In this scenario there is a strong demand for word form generators or morphologisers to assist the lookup of variant forms.

The ‘Follow references’ scenario (see above) is handled

through better development tool support, in particular Javascript development and debugging is notoriously difficult and time consuming and (3) to easily create and maintain a codebase where web-based and desktop versions can be created using the same codebase by only specifying different compilation targets.

We have thus created an architecture which supports both online and offline communication between the LAT annotation and exploration tools using Flex technology. The LAT Bridge can be used in different tools configurations in both local and remote scenarios. The LAT Bridge automatically checks at constant intervals if a network connection to our servers is available and depending on the availability of a network connection it can automatically switch between online and offline mode. When switching to online mode the LAT Bridge automatically synchronizes locally created or modified language data files with the corresponding files on the server.

The current LAT Bridge has been developed as an AIR[2]-based desktop application with clearly defined APIs for interaction and data exchange. We decided to develop the LAT Bridge as a standalone tool (instead of just adding LAT Bridge functionality to each of the LAT tools involved) to provide a flexible communication scenario where all communications are handled by a single component. Using this approach, the LAT tools do not need to "know" how to connect to other LAT tools or which LAT Tool is used in the interaction. Requests to other LAT tools are sent to the LAT Bridge in a generic way (such as "give me the list of available lexicons") and the LAT Bridge then handles the request by making an API call to the appropriate tool.

The offline communication scenario makes use of (1) Merapi[3] (a Java-AIR bridge) for communication with ELAN and (2) the Flash Player-based LocalConnection class to interact with the desktop and web based versions of ANNEX and LEXUS. The LocalConnection class allows any number of Flash Player-based applications (both AIR- and browser-based) running on the same computer to directly communicate with each other without an Internet connection or any other specific setup.

The online communication scenario is largely based on the use of web services using the services WSDL files. This currently limits the use of the LAT Bridge to only interacting with SOAP services, but poses no significant limitations in our current LAT Tool suite setup. If the web services become temporarily not available, then the LAT Bridge can easily switch to a local communication scenario. In this local communication scenario, the LAT Bridge directly communicates with the web-based versions of ANNEX and LEXUS without the need to launch the desktop versions of ANNEX and LEXUS.

5. Exchange format

The LAT Bridge is designed to deliver data in a uniform manner. This allows applications using the LAT Bridge to remain agnostic about the sub system being approached. Standardization of interchange formats thus is an important requirement for interoperability and the possibility to interchange information between different functionally equivalent sub systems.

For LEXUS the interchange format is based on the proposed LMF standard and largely follows the recommendations followed in the standard's proposed DTD (ISO FDIS 24613:2008). However, LEXUS allows each lexicon to express its own structure and as a system thus contains a number of heterogeneously structured lexica. This requires a number of user guided steps to

[2] Adobe Integrated Runtime: http://www.adobe.com/products/air/

[3] http://www.merapiproject.net

other types of user interactions such as bootstrapping word lists or searching for annotation examples. As a result, more LAT tools are expected to be integrated into the LAT bridge and further extensions of to the tools themselves are foreseen to accommodate for this.

7. Acknowledgements

We would like to thank the abstract reviewers for their helpful contributions to improve the quality of this paper.

8. References

ISO FDIS 24613:2008, Language resource management — Lexical markup framework (LMF).

Berck, P., Russel, A. (2006). ANNEX - a web-based Framework for Exploiting Annotated Media Resources. In *Proceedings of Language Resources and Evaluation Conference (LREC) 2006*. Genoa, Italy, pp. 5-22.