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.

Although the LAT tools have proven their usefulness for the linguistic domain they largely operate independently from each other, i.e. in most cases there is no direct data exchange between them. There is an increasing demand from the user community to be able to use different combinations of LAT tools in conjunction with each other requiring a flexible setup with direct interchange of data. Thus far researchers at our institute have used Toolbox[1] to link annotations and lexical information. Due to several limitations and workflow design weaknesses of Toolbox, our researchers have repeatedly asked for a optimised workflow which enables them to directly transfer and use linguistic data between the LAT annotation and exploration tools in order to create richer linguistic resources in a straightforward and efficient manner.

This paper presents a software module, the LAT Bridge, which enables bidirectional communication between the LAT annotation and exploration tools. We first briefly introduce the LAT tools and present a number of typical use case scenarios for interactions between them. We explain technical challenges and the software architecture designed and developed for the LAT Bridge.

2. Description of LAT annotation and exploration tools

Currently there are three main LAT tools for the annotation and exploration of linguistic data: ANNEX, LEXUS and ELAN. ELAN is a professional tool for the creation of complex annotations on video and audio streams. ANNEX is a web-based annotation exploration tool for metadata-based language archives. It supports many different annotation input formats (such as Shoebox/Toolbox, CHAT and EAF). LEXUS is a web-based tool to create and edit lexical databases. LEXUS is based upon the ISO 24613 Lexical Markup Framework (LMF) model which provides a common model for the creation and use of electronic lexical resources and to manage the exchange of data between and among these resources.

3. Interaction scenarios

The following basic interaction scenarios between the LAT annotation and exploration tools have been identified in collaboration with researchers at our institute:

Follow references:
A user finds a reference in an annotation or in a lexicon and wants to start the complementary tool to show exactly the fragment that has been meant - either a lexical entry or an annotation.
A reference in a lexicon entry for example may contain a reference to an annotation fragment, including time and duration information, which will be used to directly access the annotation information. Alternatively, an annotation may contain a reference to a lexical entry, or a part thereof, and is able to display this information using LEXUS.

[1] http://www.sil.org/computing/toolbox/

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.

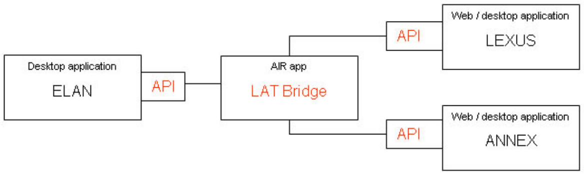
Word Completion and Correction:
A user is creating an annotation and wants to use lexical knowledge in the form of completion or correction. While the user enters the annotation information, the lexicon is automatically checked to determine whether the information is already available in the lexicon. 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 as part of the standard functionality of server based LAT tools such as ANNEX and LEXUS. These tools provide the possibility to refer to information fragments using URIs which are also commonly used by the tools themselves.
Other interaction scenarios may be realized using several LAT tools in combination with each other where each performs some of the basic functions required by the scenarios. The ‘Search for an Annotation Example’, for example, combines functionality found in LEXUS, TROVA (the archive’s search engine) and ANNEX while the ‘Word Completion and Correction’ involves interaction between ANNEX, ELAN and LEXUS. To be able to accommodate for different LAT tools to interact in a uniform manner, a bridge component was created that serves as the central communication point for all interacting tools. As an initial approach ELAN, ANNEX and LEXUS are considered to participate in the interaction process to demonstrate the technical viability of the approach. More specifically, actions such as searching for a lexical entry and basic word completion have been chosen as the primary interaction patterns since these involve lexicon lookup which is already available as a web service offered by LEXUS.

4. Software architecture

The software architecture of the LAT Bridge is guided by two main design considerations. First, it became necessary to be able to include tools that operate in different environments with ELAN as a standalone Java tool, LEXUS as a web application/web service and ANNEX which can either be used as a standalone tool or as a web application.
A further important requirement is that 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). The LAT bridge has to be able to switch dynamically between an online and an offline mode where (1) local versions of the LAT tools are used in situations where the Internet is temporarily not available and (2) connections to server based applications are made when the user is online.

The basic architecture of the LAT Bridge is based on a Facade pattern, which is a standard software engineering design pattern providing a unified interface to a set of interfaces in one or more sub systems. It is used to simplify a number of complex interactions into a single interface. It also makes it possible to provide switches in the Facade component which enables us to change the sub systems being used without requiring changes to the systems making use of the Facade component. For the LAT Bridge we can therefore provide dynamic switches between online and offline mode. In addition, the sub systems become interchangeable provided that the interaction between the LAT Bridge and the approached subsystems remains stable. Overall the standardization of interfaces becomes an important software design aspect. For applications using the LAT Bridge the standardization of information formats represents an important factor because it allows the interpretation of information obtained from potentially unknown systems.



Our technology selection is guided by a strong need to use a coherent set of technologies to allow cost effective development of the LAT tool suite. Introduction of new technological frameworks increases the level of required technical expertise and poses challenges to maintaining that expertise in the long run. For LEXUS and ANNEX the decision was made two years go to switch from a HTML/Javascript development paradigm to Flex based development (1) to improve cross-browser compatibility, (2) to gain better control over the development cycle 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 allow identification of the information elements to be used for scenarios like 'Word Completion and Correction'. These steps are reflected in the web service interface LEXUS exposes. First a lexicon needs to be selected that will be used for the lookup process. To make this selection, the LEXUS web service is able to deliver the list of lexica the current user has access to. Next, the sections in the lexicon need to be identified in which to search. For this, each lexicon maintains a set of data categories that contain the information elements for each lexicon. These may be requested through the web service interface. Query construction takes place at the client side for which an XML list of lexical entries that match the specified query are returned by the service interface. The structure of these lexical entries reflects the schema that is applicable for the selected lexicon. From this structure the user selects the information element she is interested in using, for example, in the word completion process. This will usually only need to be done once. At a later stage manual intervention by the user is only required in cases where multiple information elements are retrieved in a lexical entry, for example multiple senses, or where multiple lexical entries match the query.

The interchange format strategy used here thus allows for a large degree of flexibility where each lexicon source maintains its own structure.

6. Outlook

Currently the user of the LAT Bridge has to provide herself the lemmatised form of a text token to find the corresponding entry in an existing lexicon. Therefore we will investigate to which extent we can deploy tokenizers, morphologisers and word form generators to help researchers with this task. This is going to be a nontrivial task because our researchers often work with less commonly spoken languages (for example endangered languages) for which there are in most cases no language tools and language training data. The LAT bridge itself has proven to be useful in managing interactions between different LAT tools which gives way to further support

[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.

Kemps-Snijders, M., Nederhof, M.-J., Wittenburg, P. (2006). LEXUS, a web-based tool for manipulating lexical resources. In *Proceedings of Language Resources and Evaluation Conference (LREC) 2006*. Genoa, Italy, pp. 1862-1865.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of Language Resources and Evaluation Conference (LREC) 2006*. Genoa, Italy.