Extending NLP Tools Repositories for the Interaction with Language Data Resources Repositories

Thierry Declerck
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbruecken
Germany
declerck@dfki.de

Abstract

This short paper presents some motivations behind the organization of the ACL/EACL01 "Workshop on Sharing Tools and Resources for Research and Education", concentrating on the possible connection of Tools and Resources repositories. Taking some papers printed in this volume and the ACL Natural Language Software Registry as a basis, we outline some of the steps to be done on the side of NLP tool repositories in order to achieve this goal.

1 Introduction

The main goal of the ACL/EACL01 "Workshop on Sharing Tools and Resources for Research and Education" is to discuss methods for the improvement and extension of existing repositories. In this paper we briefly address one of the central discussion point of the workshop: how to achieve a close interlinking between NLP tools and NL resources repositories. We will base this discussion on the ACL Natural Language Software Registry (see (Declerck et al., 2000)) and some papers printed in these proceedings (see the list of papers in the bibliography).

The necessity of having repositories for NLP tools has already been clearly recognized in the past, and recently this topic has also been addressed within the broader context of a conference on Language Resources (see (Chaudiron et al., 2000) and (Declerck et al., 2000)). (Chaudiron et al., 2000) is essentially concerned with the question of identifying the NLP supply according to its different uses, and thus is describing a user-oriented approach to NLP tools repositories. (Declerck et al., 2000) is mainly describing the functionalities of the new version of the ACL Natural Language Software Registry, also showing how this version can overcome some of the practical problems encountered by former repositories (a summarized presentation of the ACL Registry is given below in section 2). Both papers are also discussing the problem of proposing a good taxonomy of NLP tools: user oriented versus developer oriented, top-down versus bottom-up approach, coarse-grained versus fine-grained classification and the way those classification strategies could cooperate. So for sure there is also still a need for establishing a cooperation between distinct approaches to NLP tools classification and their implementation, and a corresponding discussion is going on.

But since NLP tools are of interest only if they have *language data* they can process and transform, and Language Data Resources are only of interest if there is a clear indication on how they can be accessed and processed, there is a also a real need of establishing descriptive links between the two types of repositories in which tools on the one side and language data resources on the other side are included. This will allow people using a certain tool to easily find the type of language data they need. And the other way round: people having language data can easily find the type of tools that can produce some added-value for their data. The successful establishment of such a connection between these two types of repositories will probably require as well a partial reorganization of the NLP repositories on the one hand and the language data repositories on the other hand in order to maximally respond to the overall requirement of what at the end will be an *infrastructure*[1] for discovering, accessing and combining language related resources and tools.

This paper is specially addressing some of the extensions the ACL Registry is undergoing in order to offer a valuable contribution to this infrastructure.

2 The ACL Natural Language Software Registry

The Natural Language Software Registry (NLSR) is a concise summary of the capabilities and sources of a large amount of natural language processing (NLP) software available to the NLP community.[2] It comprises academic, commercial and proprietary software with specifications and terms on which it can be acquired clearly indicated.

The visitor of the NLSR has two types of access to the information stored in the NLSR: browsing through the hierarchically organized list of products (the maximal depth for browsing is level 3) or by querying for the specifications of the products as they are listed in the Registry. This querying functionality is helping the visitor in finding potential relevant software, since he or she is be able to formulate standard queries, whereas a menu allows to constrain the search to certain aspects of the listed products. So it is possible to query for example for all freely available morphological analyzer for Spanish running on a specific platform. Products can be listed in distinct sections. In order to know in which sections a product is to be found, the user can submit a standard query to the Registry Database.

The underlying classification of the actual version of the ACL Registry is largely based on the book (Varile and Zampolli, 1996). But this taxonomy will probably have to be further specialized and extended in order to satisfy the majority of the visitors of the NLSR. Therefore the classification can be enriched by the products submitted and/or by comments made by the visitors, introducing thus a bottom-up, developer and/or user oriented classification.

A general goal of the most recent editions of the NLSR was the simplification of the registration procedure, providing a short form to be filled by the customer. We do not request anymore an exhaustive description of the submitted product, but concentrate on few points providing a guiding for the visitor, who will have to consult the home page of the institutions or authors having submitted their product for getting more detailed information. In accordance with this simplification of the registration procedure, institutes or companies submitting their NLP products to the ACL Natural Language Software Registry are required to give their URL.

3 Extending the ACL Natural Language Software Registry

The ACL Registry was till recently a closed world, in the sense that information encoded in it could be accessed only by browsing or querying within its web page. Obviously there is a need for getting access to this information without having to activate a web browser. Therefore it was planned to provide for an XML export, since XML is the standard for exchanging structured documents. And this need was getting even more urgent after the Registry Team was asked for permission of harvesting the ACL repository for the purpose of creating a prototype service provider in the context of an Open Archive Initiative for Language Resources, which is called OLAC (Open Language Archives Community) and described in (Bird and Simons, 2001).

This excellent initiative also requires that the information provided by tools repositories is not only universally available but also has to conform to certain standards for *metadata* description. This in order to ensure the *interoperability* across all the repositories participating as metadata providers in OLAC.

4 XML for Tools Repositories

(Erjavec and Váradi, 2001) are proposing a very interesting description of the TELRI-II concerted action for a tool catalogue specialized for corpus processing tools. This “limitation” in the

[1] As (Bird and Simons, 2001) names it.
[2] See http://registry.dfki.de/

coverage of the repository TELRI repository is allowing the authors to make extensive experiments with various XML specifications and tools for the building and display of their catalogue. An experience which should be beneficial for the more generic ACL Registry, as well as for other provider of tools repositories (so for example national initiatives, like the one described in (Chaudiron et al., 2000)). The authors also mention one advantage of the limitation in the coverage of tools: the presence in the entries of a pointer to persons or institutions being able to offer advice on installing and using the software. Thus addressing also one point mentioned in (Bird and Simons, 2001), where 3 main classes of providers are described: DATA, TOOLS and ADVICE providers.

But (Erjavec and Váradi, 2001) are not proposing a discussion on how to integrate in the description of the tools the particular relation to a specific corpus. Nevertheless this should be a common task to be tackled by all providers of tools repositories. Probably it would be the best strategy to start with specialized repositories, where the problems to solve can appear earlier.

5 Metadata for NLP Tools

As we saw above, the sole conformance to standards (XML) for document description and interchange is not enough in the context of OLAC. But the use of metadata descriptions for tools seems to make sense not only for such initiatives. (Lavelli et al., 2001) show the use of metadata description for tools in the context of an infrastructure for NLP application development. The role of metadata there is to specify the "level of analysis accomplished by the source processor". Thus the metadata descriptions are useful for the communication between processes within an NLP chain, and also allow to mark and identify the document produced by such a process. In any cases, the use of metadata description for tools (or processes triggered by those tools) is probably a key-issue in the modular design of complex NLP environment.

And one can see in the SiSSA approach to metadata descriptions for NLP processes, maybe as a side effect, a proposition for sharing annotations for processes and documents (resources) that can be handled. This might be a starting point for the systematic connection of the descriptions of both NLP tools and language resources.

6 Connection with Metadata-Descriptions for (Multimedia/Mltimodal) Language Resources

Catalogue and repositories for Natural Language data resources have already been working on the topic of metadata description for their entries (See for example LDC and ELRA). One can see OLAC as a natural extension of the LDC, enlarging the resources catalogue to a real infrastructure for language resource identification.

From the side of the Language Engineering there are initiatives for describing standards and (Calzolari et al., 2001) present such an initiative, the ISLE project, which is the continuation of the EAGLES initiative. The main objective of ISLE is to promote "widely agreed and urgently demanded standards and guidelines for infrastructural language resources ..., tools that exploit them and LE products". The ongoing discussions within this project are thus important for the intended extension of NLP tools repositories.

While (Calzolari et al., 2001) concentrate on the description of the task of the ISLE computational lexicon working group and address the topic of metadata for encoding multilingual lexical resources, (Broeder and Wittenburg, 2001) presents the work of the ISLE Metadata initiative (IMDI), which is directly relevant for the topic addressed here. (Broeder and Wittenburg, 2001) give a good overview of metadata initiatives for Language Resources and propose a contrastive description of OLAC and IMDI, where the main distinction can be seen in the top-down versus bottom-up approach. The top-down approach followed by OLAC allows an easy conformance to the Dublin Core set, whereas the bottow-up approach requires the definition of more "narrow and specialized categorization schemes".

This distinction is important for the intended extension of the metadata description for NLP tools, since the description of the tools will have to connect to those distinct kinds of categorization schemes for data resources. We think here that the ACL Registry can easily be adapted to this situation since the actual classification of tools is a layered one, one layer being quite general (classifying tools wrt broader application types, like "Written Language"), and the next layer stressing more the specific technology (for example Information Extraction versus Text Alignment).

(Broeder and Wittenburg, 2001) is also proposing a scheme for connecting the descriptions of tools and resources. They suggest not to include a listing of tools in the metadata description of the resources, since this set of tools would be changing in time. Rather they suggest a detailed description of the type and the structure of the resources that can be accessed by a "browser" tool, which on the basis of the detailed metadata description can select potential tools for handling the resources. The tools repository would have to include this kind of information in its metadata description of the tools.

7 Conclusion

As we could see out of this (not exhaustive) selection of papers submitted to the ACL/EACL01 "Workshop on Sharing Tools and Resources for Research and Education", there are a lot of very interesting and promising, implicit or explicit, suggestions for the goal of connecting tools and resources repositories. The ACL Natural Language Registry will take these suggestions as the basis of the further work on providing extensions to metadata descriptions in order to be as compliant as possible to emerging infrastructures and standards for language resources.

References

S. Bird and G. Simons. 2001. The OLAC Metadata Set and Controlled Vocabularies. In *This volume*.

D. Broeder and P. Wittenburg. 2001. Interaction of Tools and Metada-Descriptions for Multimedia Language Resources. In *This volume*.

N. Calzolari, A. Lenci, and A. Zampolli. 2001. International Standards for Multilingual Resource Sharing: The ISLE Computational Lexicon Working Group. In *This volume*.

S. Chaudiron, K. Choukri, A. Mance, and V. Mapelli. 2000. For a repository of NLP tools. In *LREC 00*, pages 1273–1278.

T. Declerck, A.W. Jachmann, and H. Uszkoreit. 2000. The new Edition of the Natural Language Software Registry (an initiative of ACL hosted at DFKI). In *LREC 00*, pages 1129–1132. http://registry.dfki.de.

T. Erjavec and T. Váradi. 2001. The TELRI tool catalogue: structure and prospect. In *This volume*.

A. Lavelli, F. Pianesi, E. Maci, I. Prodanof, L. Dini, and G. Mazzini. 2001. SiSSA – An Infrastructure for NLP Application Development. In *This volume*.

G.B. Varile and A. Zampolli. 1996. Survey of the State of the Art in Human Language Technology. http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html.