The new Edition of the Natural Language Software Registry (an initiative of ACL hosted at DFKI)

Thierry Declerck, Alexander Werner Jachmann, Hans Uszkoreit

DFKI GmbH (German Research Center for Artificial Intelligence)
Language Technology Lab
Stulsatzenhausweg 3, 66123 Saarbrücken, Germany
{declerck, heinz, hansu}@dfki.de

Abstract

In this paper we present the new version (4th edition) of the Natural Language Software Registry (NLSR), an initiative of the Association for Computational Linguistics (ACL) hosted at DFKI in Saarbrücken. We give a brief overview of the history of this repository for Natural Language Processing (NLP) software, list some related works and go into the details of the design and the implementation of the new edition.

1. Introduction

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. It comprises academic, commercial and proprietary software with specifications and terms on which it can be acquired clearly indicated. The NLSR proposes a structured listing and descriptions of available NLP products, but does *not* include a distribution facility.

While the NLSR concentrates on listing NLP software, it does not exclude the listing of Natural Language Resources (NLR), since we would also like to include resources which are strongly related to the processing tools listed in the Registry. But, in general, other institutions or projects provide exhaustive listings of such resources and/or distribute them (see ELRA/ELDA or LDC, to mention but a few) and are far more competent on this topic than we could ever be. A cooperation between NLSR and those institutions here is far more indicated as duplicating entries on Language Resources at the NLSR site.

1.1. A Brief History of the NLSR

The original concept of the NLSR is due to Jessie Pinkham. The second edition, supervised by Elizabeth Hinkelman, owed much to the participants of the 1992 survey of natural language processing software conducted for the German Ministry for Research and Technology (grant ITW 9002 0 to DFKI inc.) by the DFKI and directed by Prof. Wolfgang Wahlster. The third and fourth editions have been produced by the Language Technology Lab. of the DFKI inc. under the direction of Prof. Hans Uszkoreit. With the third edition the NLSR looked forward to cooperation with initiatives and projects of the European Community, such as ELSNet (http://elsnet.let.uu.nl/resources.html), RELATOR (http://www.linglink.lu/le/projects/relator/index.html), and the software survey conducted by the University of Pisa. The 4th edition also includes suggestions made during some meetings about the integration of language resources[1].

1.2. Other Resources

Complementary information to the one listed in the NLSR concern the language resources. Interesting listings of natural language resources include:

- ACL NLP/CL Universe Resources listing (ACL: http://www.cs.columbia.edu/∼radev/u/db/acl/html /RESOURCES/)
- Bavarian Archive for Speech Signals (BAS: http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html)
- Center for Lexical Research (http://crl.nmsu.edu/cgi-bin/Tools/CLR/clrcat#TOC)
- European Language Resources Association (ELRA/ELDA: http://www.icp.inpg.fr/ELRA/)
- Linguistic Data Consortium (LDC: http://www.ldc.upenn.edu)
- Summer Institute of Linguistics: Linguistic Resources on the Internet (SIL: http://www.sil.org/linguistics/topical.html)

There are also national listings of NLP software, with which we will look for a closer collaboration, an example being the “Répertoire d'outils pour le Traitement Automatique des Langues” (http://www.biomath.jussieu.fr/ATALA/outil/) in France.

2. The Fourth Edition of the NLSR

We present here the fourth edition of the NLSR, which offers, among others things, a new functionality: a database has been developed which allows menu-guided standard queries for all the specifications of the products as they are listed in the Registry. So the visitor will have two types of access to the information stored in the NLSR: browsing

[1]We would like to thank the organizers (Univ. of Sheffield) and the participants of the Baslow Workshop on NLP Architectures and Language Resources (see http://www.dcs.shef.ac.uk/∼hamish/dalr/baslow/). Our thank also to the ELRA Team for the one day meeting they organized at their site.

through the structured list of products or by asking for specific information.

Another novelty concerns the classification of the products, which in the former editions of the NLSR no longer accurately represented the state of the art of Human Language Technology (HLT). To establish such a taxonomy is a far from trivial task, and the classification proposed here will probably have to be further specialized and extended in order to satisfy the majority of the visitors of the NLSR. Our taxonomy is largely based on the book (Varile and Zampolli, 1996). The classification can be enriched by the products submitted and/or by comments made by the visitors.

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.

Also the work of the administrators of the NLSR has been now greatly simplified through the implementation of a "control and editing center". Thus the communication between the Registry Team and the submitters has been made more efficient and transparent. Additionally an updating facility will be provided very soon.

The 4th version will also control whether the software listed in the registry is kept up-to-date, thus avoiding the presence of no longer maintained products.

2.1. The new Classification of the NLP Software

The NLP software products submitted to the Natural Language Software Registry are now hierarchically organized, and the user can browse through the proposed classification. The maximal depth for browsing is level 3. Some of the products are 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.

As already stated in section 2., the adopted taxonomy is largely based on the book (Varile and Zampolli, 1996). Links are provided for the sections having a direct correlation to chapters of this book. This information will also help the submitters to decide on the sections they would like to see their products be listed in. We are confident that we will get a still more accurate taxonomy, depending on the number and the kind of submissions we receive, and also on the base of comments the visitors might propose on this topic. We will also additionally consult other overviews on the state of the art of NLP, like (NSF/EU, 1999).

The (partial) listing below will give an overview of the actual implemented taxonomy. In some case we also show the further sub-classification (level 3 is indicated in brackets):

- Spoken Language
  - Spoken Language Input (Speech Recognition, Speaker Recognition, Spoken Language Understanding)
  - Spoken Language Output (Speech Synthesis, Spoken Language Generation, Text-to-Speech Synthesis)
  - Spoken Dialog Systems
  - Spoken Language Translation

- Written Language
  - Document Processing (Optical Character Recognition (OCR), Document Image Analysis, Tokenization, Processing Mark-Up Languages, Information Retrieval, Information Extraction, Summarization, Document classification, Handling Controlled Languages, Text Mining, Terminology Extraction)
  - Analysis (Morphological Analysis, Part-of-Speech Tagging, Shallow Parsing, Partial Parsing, Deep Syntactic Analysis, Semantic & Pragmatic Analysis)
  - Language Generation (Shallow Generation, Deep Generation)
  - Written Language Translation
  - Written Dialog Systems
  - Authoring Aids

- Multi-modality
  - Text and Images
  - Speech and Gesture
  - Facial Movement & Speech Recognition
  - Facial Movement & Speech Synthesis

- Language Resources (only those related to the tools listed)
  - Written Language Corpora
  - Spoken Language Corpora
  - Multi-modal Corpora
  - Lexicons
  - Terminology
  - Formalisms

- Multi-media Systems
  ...

- Evaluation Tools
  ...

- Annotation Tools
  ...

- Others
  - Knowledge Representation Tools (?)
  ...

2.2. Asking the Database of the NLSR

This new functionality will help the visitor in finding potential relevant software, since he or she will be able to formulate standard queries and a menu will allow to constrain the search to certain aspects of the listed products. So it is now possible to query for example for all freely available morphological analyzer for Spanish running on a specific platform. The system shows all the results of the query and the user can immediately have access to those since the answers also include the URL of the institution or company submitting the software. The Querying Interface is shown in table 1.

2.3. The Submission Procedure

All authors of systems dealing with natural language are invited to contribute to the ACL Natural Language Software Registry, using for this the submission form displayed at the address given in section 2.4. This Form is also (partly) shown in table 2. Some of the fields of this form must be completed (”URL”, ”Name”, ”Author(s)”, ”Affiliation”, ”Description”, ”Abstract”, ”Mail”). The system checks if all the mandatory slots of the submitted form have been completed: and if not, it presents the user with a list of the non-filled mandatory fields, without leading back to the whole submission form. Some fields require text-input, and in other cases the user can just select one or more of the predefined values in a field (more than one choice is possible).

The information you put or select in the various fields will be stored in the NLSR database and can later be accessed by standard queries.

If some of your specifications are not present in the list of predefined values for a field, this information can be given in the field ”others” of the form. The Registry Team will then add the missing values to the set of predefined values for this case, thus continually extending and updating the submission form.

2.4. URL of the NLSR

- the third edition – http://www.dfki.de/lt/registry/
- the fourth edition – http://registry.dfki.de

Both versions will co-exist a limited amount of time for browsing, until all the entries of the 3rd version have been checked and ported to the new version. For submitting new products, only the fourth version will be available.

3. Future Work

Additionally to the ongoing improvement of the proposed taxonomy and the controlled increasing of the number of entries, we would like to investigate the role that an initiative like the NLSR can play in the future development of web-based NLP architecture, and thus see how it can contribute to the success of Human Language Technology applications. In a sense we think that such initiatives should be in the future more involved in the kind of discussions which have already been successfully conducted about the integration of language resources. A promising beginning of such a discussion has been offered at the “Baslow Workshop on NLP Architectures and Language Resources” and we will try to continue elaborating on the base of those results.

4. Acknowledgments

We would like to thank Jasmin Schneider for her help in designing the Registry home page, which is based on an illustration of the first calculating machine (1623), developed by Wilhelm Schickard in Tübingen (see http://www-groups.dcs.st-and.ac.uk/∼history/Mathematicians/Schickard.html).

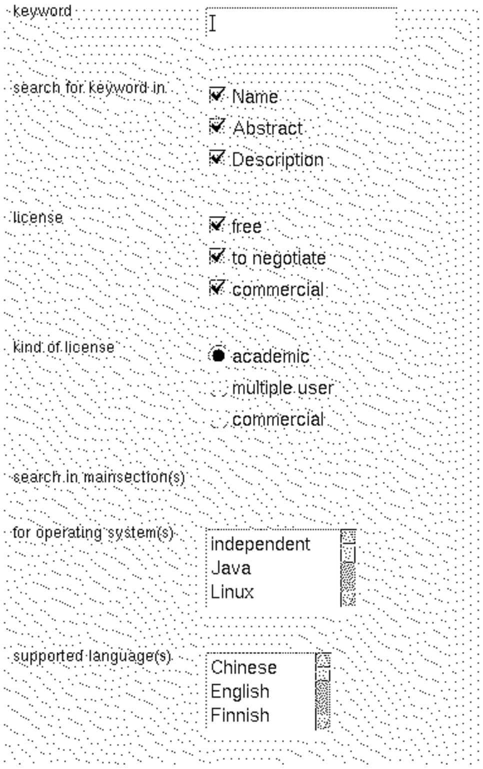
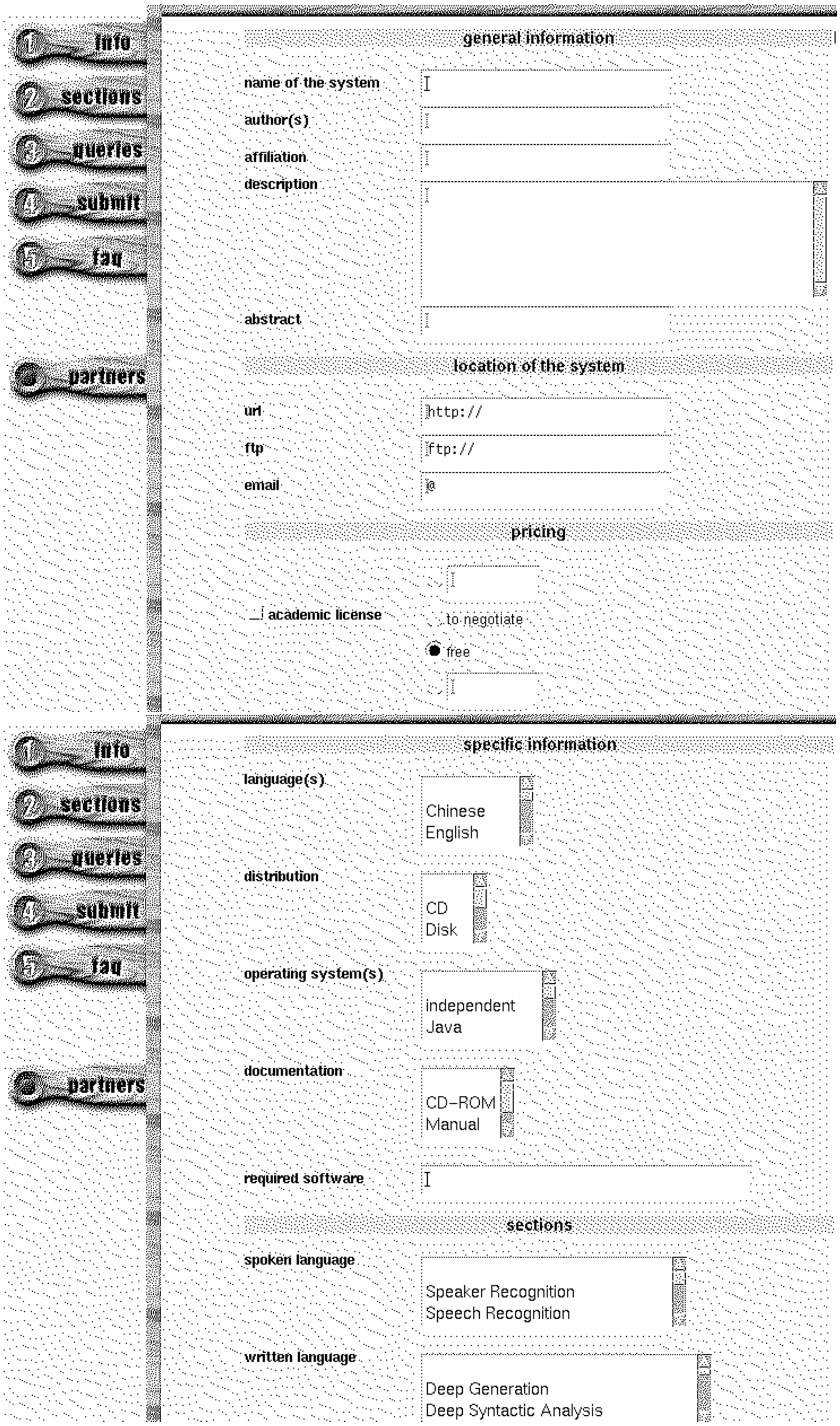


Table 1: The Querying Interface of the NLSR

5. References

NSF/EU, 1999. Multilingual information management: Current levels and future abilitiesbibliographic references. Technical report, US National Science Foundation. Also available as: http://www.cs.cmu.edu/∼ref/mlim/index.html.

Varile, G.B. and A. Zampolli (eds.), 1996. *Survey of the State of the Art in Human Language Technology*, volume XII - XIII of *Linguistica Computazionale*. Pisa. Also available as: http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html.



1 info
2 sections
3 queries
4 submit
5 faq
partners

general information

name of the system
author(s)
affiliation
description
abstract

location of the system

url http://
ftp ftp://
email @

pricing

academic license
to negotiate
free

1 info
2 sections
3 queries
4 submit
5 faq
partners

specific information

language(s) Chinese English
distribution CD Disk
operating system(s) independent Java
documentation CD–ROM Manual
required software

sections

spoken language Speaker Recognition Speech Recognition
written language Deep Generation Deep Syntactic Analysis

Table 2: The Submission Form of the NLSR