Language Technology Resource Center

Jennifer DeCamp
MITRE Corporation
7515 Colshire Drive
McLean, VA 22012, USA
jdecamp@mitre.org

Abstract

This paper describes the Language Technology Resource Center (LTRC), a U.S. Government website for providing information and tools for users of languages (e.g., translators, analysts, systems administrators, researchers, developers, etc.) The LTRC provides information on a broad range of products and tools, and provides a means for product developers and researchers to provide the U.S. Government and the public with information about their work.

1 Background

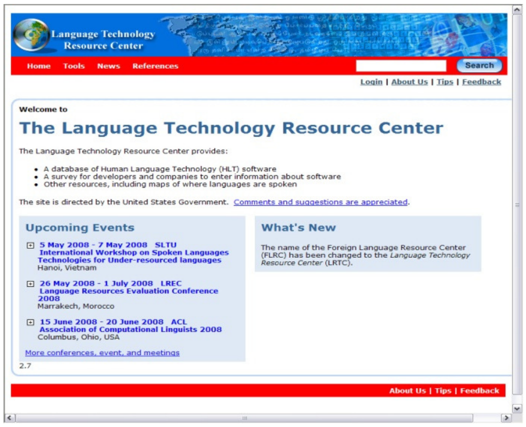
The Language Technology Resource Center (LTRC) is a U.S. Government website that provides information on a broad range of products and tools. It also includes an online survey for product developers and researchers to provide the U.S. Government and the public with information about their work.

The LTRC is developed and run by MITRE. As described on MITRE's home page, "The MITRE Corporation is a not-for-profit organization chartered to work in the public interest. As a national resource, we apply our expertise in systems engineering, information technology, operational concepts, and enterprise modernization" to address U.S. Government needs.

2 History

In 2000, in response to the crisis in Albania, MITRE began posting and distributing information about available tools to support the relief effort. In 2002, the effort received a small amount of funding from the Office of the Secretary of Defense. In 2004, it became one of the first projects of the Language and Speech Exploitation Resources (LASER) Advanced Concept Technology Demonstration (ACTD).

The program was transitioned to the U.S. Army and currently reports to the Foreign Language Program Office in the Army G2. It was transitioned to the Defense Intelligence Agency Foreign Language Management Division in 2009.


Language Technology Resource Center
Home
Tools
News
References
Search
Login | About Us | Tips | Feedback
About Us | Tips | Feedback

Welcome to

The Language Technology Resource Center

The Language Technology Resource Center provides:

- A database of Human Language Technology (HLT) software
- A survey for developers and companies to enter information about software
- Other resources, including maps of where languages are spoken

The site is directed by the United States Government. Comments and suggestions are appreciated.

Upcoming Events

⊞ 5 May 2008 - 7 May 2008 SLTU
International Workshop on Spoken Languages Technologies for Under-resourced languages
Hanoi, Vietnam

⊞ 26 May 2008 - 1 July 2008 LREC
Language Resources Evaluation Conference 2008
Marrakech, Morocco

⊞ 15 June 2008 - 20 June 2008 ACL
Association of Computational Linguists 2008
Columbus, Ohio, USA

More conferences, event, and meetings

What's New

The name of the Foreign Language Resource Center (FLRC) has been changed to the *Language Technology Resource Center* (LRTC).

2.7

Figure 1. Screen Capture of LTRC.

In 2007, the name of the website was changed from the "Foreign Language Resource Center" to the "Language Technology Resource Center" in order to better reflect the content. In 2009, the LTRC was moved to the Defense Intelligence Agency (DIA) Language Technology Management Office.

3 Mission

The mission of the LTRC and the FLRC, as defined in the website, is to obtain and share information about language-related software that is or soon will be available from academia, industry, government, or other sources.

4 Requirements

The requirements of the LTRC and FLRC, as described in the 2005 Management Plan, are to:

4.1 Identify Language Tools

Requirements for information on language tools were collected from U.S. government translators, translation managers, researchers, language teachers, users of searches and translations, and systems administrators. Information needs were included from four ongoing or planned Government technology surveys and from early stages of the Framework for Evaluation of Machine Translation in ISLE (FEMTI). ISLE was a European Community program for International Standards for Language Engineering.

Prior surveys had shown that sufficient information was not included in many product websites and/or that such information was buried in the site where it could not be extracted easily with automated methods.

4.2 Facilitate Acquisition of Tools

Requirements were also collected from acquisitions officers, whose needs included a listing of competitive products, information on compliance with Section 508 on accessibility, and availability of the product on the Government Services Administration (GSA) Schedule.

4.3 Plan

In addition, planners needed to assess language readiness, which was challenging considering the number of languages and dialects, the number of types of technology, and frequently the number of products per language/dialect per type.

5 Implementation

The website is implemented with the following:

5.1 Tool Survey

For The site includes an online tools survey based on information needs from government planners, acquisitions officers, systems administrators, programmers, and users. The only required information is the name of the company and product, a short description of the product, and the name and contact information for further information. However, people are welcome to complete as much of the form as they wish, as well as to send electronic information such as white papers and evaluations.

People completing the survey are contacted each six months with reminders to update information. The FLRC team also keeps track of new product announcement and solicits updates of the survey.

Information can be entered display on the public internet and/or for display on government networks.

5.2 Tool Database and Reports

Reports can be generated from the tools database. The most common report is a particular language crossed with a particular type of technology (e.g., Arabic x Optical Character Recognition [OCR]). A user can also select a language and see all technologies, or select a technology and see all languages.

Another type of report is a matrix of languages and types of tools. Each cell includes the number of products of a particular type of technology in the database that work with that language (e.g., the number of Chinese OCR products). The original plan had been to develop metrics and measures whereby a cell could be colored red, yellow, or green to show language readiness. However, agreement has not been reached on these measures.

A separate search capability is provided for dictionaries in order to capture the many types of tools available. A standard dictionary search includes dictionary type, source language, target language, domain, register, and year of publication. An advanced search includes the ability to further define these parameters.

A software comparison capability is also provided (in prototype), where users can select the type of technology and the language or dialect. They could then select from a list of requirements which are hard (non-negotiable), soft (desired), or null (no requirement). The system then ranks the products in the database according to users' selections.

Information on cost and accuracy is provided separately, as this information is relative. For instance, an additional $200 may be well worth two percentage points in accuracy. One possibility that has been considered is to enable users to select cost and accuracy on a sliding scale.

The issue of accuracy, of course, is still in debate and in development for many types of language technology tools. A possibility in the future may be to list test scores along with the tasks that the system may be able to do. Additional research is needed for such measures.

A further report is provided on future products in a particular language and technology type. This information is important for many organizations, as it may influence decisions to acquire software or to wait for the new products. However, due to the company sensitivity of the information, there are few entries in this area and none open to the general public.

5.3 Other Software Tools

The tools section of the LTRC includes links to online MT. It also includes downloadable virtual keyboards for Chechen, Dari, Dinka, Iraqi Arabic, Kurdish, Pashtu, Tajik, Turkmen, Uighur, Urdu and other languages. The keyboards work with Microsoft Windows 2000 and higher.

5.4 FAQs

FAQs are provided for Arabic-script languages. There is potential for linking to many other FAQs available from Microsoft and from other sources on the internet.

5.5 Information

Information is provided on the following:

Conferences and Meetings
Information is provided on conferences and meetings. Information can be provided by anyone with a password throughout the government community.

Announcements
Information is provided on announcements. Information can be provided by anyone with a password throughout the government community.

Standards
Information is provided on standards from ISO, the Institute of Electrical and Electronics Engineers (IEEE), the Localization Standards Association (LISA), the American Translators Association (ATA), the World Wide Web Consortium (W3C), the Object Management Group (OMG), the International Standards for Language Engineering (ISLE), and other organizations related to language. Where permitted by copyright, the standards or links to downloadable standards are included. Standards in development are also included.

Organizations
Information is provided on professional organizations and on U.S. government organizations. Information can be updated by any person with a password via an internet form.

Resources
Information is provided on a wide range of resources, including&. Information can be updated by any person with a password via an internet form.

6 Related Efforts

There are several related efforts to catalog tools and resources. Some of the major ones are listed below.

6.1 Compendium for Translation Software

John Hutchins maintains the Compendium for Translation Software for the European Association for Machine Translation (EAMT) and the International Association for Machine Translation (IAMT). The Compendium catalogs the following:

Mt Systems
MT systems include: “for home use (e.g. for personal use by the general public); for Internet/Web (i.e. for translating electronic documents on the Internet, electronic mail, Web pages, chat discussions, etc.); for professional use (e.g. for use by professional translators); for company intranets.”

Translation Support Tools
Translation support tools include: “Electronic dictionaries; Localization support tools; Translation memory systems; Alignment tools; Terminology management systems; Foreign language authoring systems; Translator workstations.”

Online Systems
Online systems include: “MT services (i.e. translation service via Internet (or mobile telephone), with or without human post-editing); MT portals (i.e. services on Internet providing access to a number of MT services and/or to information about MT systems).”

Language pairs and contact information are provided. The Compendium is available in PDF form from the EAMT website.

6.2 Association for Computational Linguistics Natural Language Software Registry

The Association for Computational Linguistics (ACL) 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." It is oriented towards research.

6.3 European Community's Common Language Resources and Technology Infrastructure

The European Commission's Common Language Resources and Technology Infrastructure (CLARIN) is a new program as of 2008 "aiming at uniting existing digital archives in Europe that contain language-based material into a federation that will allow the social sciences and humanities research communities unified access to the content." It will make speech and text tools "available to interested researchers with a view to opening up new research avenues." It will also "provide web based services that will allow non-expert users (especially humanities and social sciences researchers without a technological background) to perform complex tasks on the materials contained in the archives." CLARIN is oriented towards European languages and towards research.

6.4 MEDAR

The European Commission's Mediterranean Arabic Resources (MEDAR) also began in 2008. It has the objectives of:

"Consolidating a network of players in all areas of HLT

Developing the Cooperation Roadmap based on a clear picture of the foreseeable technological trends, market potentials, and cooperation possibilities

Updating the Basic Language Resource Kit: the minimum set of resources and tools necessary for carrying out research and training on LRs [Language Resources] and HLT [Human Language Technology], with a focus on MT and MLIR [Multilingual Information Retrieval] Supporting the development of tools and resources, in particular MT and MLIR on the basis of partners' technologies and open source code (e.g. Statistical MT, MLIR, and speech recognition) and the framework for their benchmarking".

6.5 FLaReNeT

FLaReNet is a European Commission eContentPlus project started in 2008. It is intended to "promote the consolidation of a European strategy in the field of Language Resources and Language Technologies and to be a European forum to facilitate interaction among LR stakeholders." Thematic areas and working groups include:

The Chart for the area of LRs [Language Resources] and LT [Language Technology] in its different dimensions.

Methods and models for LR building, reuse, interlinking, maintenance, sharing, distribution

Harmonization of formats and standards

Definition of evaluation and validation protocols and procedures

Methods for the automatic construction and processing of LRs.

6.6 Other

The European Language Resources Association (ELRA) maintains a catalog of language resources, which are primarily corpora and lexicons. The Linguistic Data Consortium (LDC) is similarly focused on corpora. There are also various lists of tools on the internet, usually arranged by language and/or by type of tool (e.g. http://www.yourdictionary.com).

7 Issues

Issues include: terminology; separation and aggregation; varying requirement sets; sensitivity of information; completeness; accuracy; currency; preparedness; context; and leveraging other resources.

7.1 Terminology

One problem in developing a survey was the disagreement of what constitute some technology. For instance, several machine translation companies (including rule-based MT companies) that at the time had no tools for translators checked the box for "Translation Management System". While MT can be thought of as a system for managing translation, the type of tool known as a "Translation Management System" is one directed towards human translators, providing Translation Memory and support tools.

There are currently several efforts to standardize terminology, including work in the International Organization for Standardization (ISO) Technical Committee (TC) 37, Subcommittee (SC) 4 on Language Resources Management.

7.2 Separation and Aggregation

During the last few years, there have been the dual trends of separation and aggregation. Tools such as morphological analyzers were being marketed by themselves to be incorporated into a wide range of products. At the same time, many tools were being aggregated, particularly with MT. For instance, the Army Research Laboratory integrated OCR and MT for several language pairs. The combination then was used in many larger systems.

7.3 Wide Range of Requirement Sets

One of the difficulties in providing information is that the U.S. Government has a wide range of requirement sets. One tool or one configuration may be recommended for one application but not for another. See Section 5.2 on comparison reports.

7.4 Sensitivity of Information

Information is often sensitive—particularly concerning companies' plans and product evaluations. Some companies enter data only to be displayed on government networks and not on the public internet. In some cases, contact information for such data. In other cases, the information is just sent to government parties as approved by the data owner.

7.5 Completeness

The usefulness of the LTRC depends on information that covers at least the top several products in each technology type and language. Coverage has been an issue, with some companies eager to provide data and some not. In some cases, MITRE has provided the information, with a notation in the description field that the information was provided by MITRE and not by the owner of the information (i.e., the developer, company, or researcher). Reviews for completeness are conducted by Subject Matter Experts in MITRE and other organizations. In addition, users of the LTRC provide suggestions.

7.6 Accuracy

Accuracy is also critical to the LTRC. Currently, entries for tools are reviewed by the LTRC team. Once approved, the entry displays on the website. When a tool is evaluated and all items in the product entry are validated, a green "V" (for Validated) appears next to the product name wherever it appears in an LTRC report.

The approval review was needed in the earlier years of the LTRC, as there were frequent entries from companies announcing "99% accuracy" in machine translation and other difficult-to-substantiate claims. Due to the increasing maturity of the language technology field, such entries are now scarce. The approval stage may soon be eliminated so that entries appear as soon as the person completing the survey hits the "submit" button.

Users of the LTRC almost always want to know levels of accuracy and usefulness. Where possible, the LTRC provides government users with existing evaluations. The LTRC also runs evaluations of systems using the target data.

In the product survey, there is a section for data owners to add measures and metrics, along with relevant data such as the date and conductors of the evaluation and a link (if possible) to the evaluation. However, due to the sensitivity of many of these evaluations, few data owners have provided this information in the web survey.

7.7 Currency

Currency of data is also critical. Some companies are prompt in providing data on any product updates. Some are not. In some cases, as described previously, MITRE defaults to writing the entry themselves (with the entry so designated), with an offer to replace the entry with data provided from the company or developer.

Administrative tools track the status of product entries, including the dates when the entries were last updated. Reminders are sent to data Points of Contact (i.e., those people who filled out the surveys) every six months to remind them to update their information.

Administrative tools track the status of product entries, including the dates when the entries were last updated. Reminders are sent to data Points of Contact (i.e., those people who filled out the

surveys) every six months to remind them to update their information.

7.8 Preparedness

A key concern is to be prepared for U.S. Government language requirements, particularly since there is often little or no lead time between a crisis and the requirement for available, tailored, integrated products. A particular challenge is the area of disaster relief, where there is a high need for language interaction (e.g., medical questions, refugee registration, etc.) but little or no notice.

7.9 Context

Another concern is to provide context for the product information sheets. LTRC users not familiar with language technology need not just a list of tools (e.g., Arabic<->English MT) but also information on when they will also need OCR, OCR clean-up tools, search, or other tools and which of these additional tools have already been integrated with the MT. This concern is now being addressed with reports and guides which link to the product information sheets.

7.10 Leveraging Other Resources

With the rapidly increasing amount of information and the small amount of funding for language technology knowledge management, there is a need to leverage national and international resources. Links are helpful but often not sufficient. Guides (as described above) may lay out a search strategy with links. Additional means of leveraging these other resources are being explored.

8 Plans

Plans include the following:

8.1 Input

The LTRC team continues to solicit companies, developers, and researchers to add information via the survey tool. In some cases, the team has provided information themselves, then emailing the link to the information to the company. That information is marked as having been entered by MITRE rather than by the owner of the information. The LTRC team is also soliciting input on the structure of the website, as changes are being made to the infrastructure and the content.

8.2 Review

The team also continues to solicit review of completeness and accuracy of information. There are plans to make better use of professional organizations for this review.

8.3 Update of Survey

The survey is being updated to reflect changing requirements and tools such as the Framework for Machine Translation in International Standards in Language Engineering (ISLE) or FEMTI.

8.4 Wikis

Wikis are planned to enable comments on products and guides.

8.5 Push Technology

There are plans for providing push technology, in order to enable users to automatically receive information about announcements, conferences, meetings, and new or updated products of interest (e.g., all products supporting Hindi).

8.6 Newsletter

The LTRC has provided articles for other newsletters. Plans include continuing this practice and perhaps also providing an LTRC newsletter.

9 Coordination

Most important is continued coordination with other efforts (e.g., CLARIN, MEDAR) in order to leverage work that is already done or that may be planned. In addition, there is a need to identify and develop means of providing users with the easiest and most complete possible access to information about language technology.

Acknowledgments

I would like to thank the developers and content providers for the LTRC, including Mark Anderson, Dave Cedel, Carsten Oertel, and Derek LaHousse. In addition, I would like to thank the United States Defense Intelligence Agency Foreign Language Program Division for their sponsorship of this project.

References

ACL Natural Language Software Registry website:

http://registry.dfki.de/

CLARIN Newsletter #1, May 2008.

ELRA Catalog of Language Resources:

http://catalog.elra.info/

FEMTI: http://www.isi.edu/natural-language/mteval/

FLRC Funding Proposal, 2006.

Hutchins, J.W. (2008). *Compendium of Translation Software directory of commercial machine translation systems and computer-aided translation support tools*, developed on behalf of EAMT and IAMT, http://www.hutchinsweb.me.uk/Compendium.htm

ISO TC37 SC4 homepage: http://www.tc37sc4.org/

Language Technology Resource Center,

http://ltrc.mitre.org.

Linguistic Data Consortium:

http://www.ldc.upenn.edu/Catalog/

MITRE: http://www.mitre.org.