The SILT and FlaReNet International Collaboration for Interoperability

Nancy Ide
Department of Computer Science
Vassar College
Poughkeepsie, New York USA
ide@cs.vassar.edu

James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, Massachusetts USA
jamesp@cs.brandeis.edu

Nicoletta Calzolari
CNR-ILC
Pisa, Italy
glottolo@ilc.cnr.it

Claudia Soria
CNR-ILC
Pisa, Italy
claudia.soria@ilc.cnr.it

Abstract

Two major projects in the U.S. and Europe have joined in a collaboration to work toward achieving interoperability among language resources. In the U.S., a project entitled ”Sustainable Interoperability for Language Technology” (SILT) has been funded by the National Science Foundation under the INTEROP program, and in Europe, FLaReNet Fostering Language Resources Network has been funded by the European Commission under the eContentPlus framework. This international collaborative effort involves members of the language processing community and others working in related areas to build consensus regarding the sharing of data and technologies for language resources and applications, to work towards interoperability of existing data, and, where possible, to promote standards for annotation and resource building. In addition to broad-based US and European participation, we are seeking the participation of colleagues in Asia. This presentation describing the projects and their goals will, we hope, serve to involve members of the community who may not have been aware of the effort before, in particular colleagues in Asia.

1 Overview

One of today’s greatest challenges is the development of language processing capabilities that will enable easy and natural access to computing facilities and information. Because natural language processing (NLP) research relies heavily on such resources to provide training data to develop language models and optimize statistical algorithms, language resources–including (usually large) collections of language data and linguistic descriptions in machine readable form, together with tools and systems (lemmatizers, parsers, summarizers, information extractors, speech recognizers, annotation development software, etc.)– are critical to this development.

Over the past two decades, the NLP community has invested substantial effort in the creation of computational lexicons and compendia of semantic information (e.g., framenets, ontologies, knowledge bases) together with language corpora annotated for all varieties of linguistic features, which comprise the central resource for current NLP research. However, the lack of a thorough, well-articulated longer-term vision for language processing research has engendered the creation of a disjointed set of language resources and tools, which exist in a wide variety of (often incompatible) formats, are often unusable with systems other than those for which they were developed, and utilize linguistic categories derived from different theoretical frameworks. Furthermore, these expensive investments are often produced only for one of several relatively isolated subfields (e.g., NLP, information retrieval, machine translation, speech processing), or even worse, for one application in one subfield. In addition, the high cost of resource development has prevented the creation of reliable, large-scale language data and annotations for many phenomena, and for languages other than English.

Interoperability of resources, tools, and frameworks has recently come to be recognized as perhaps the most pressing current need for language processing research. Interoperability is especially

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 178–181,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

critical at this time because of the widely recognized need to create and merge annotations and information at different linguistic levels in order to study interactions and interleave processing at these different levels. It has also become critical because new data and tools for emerging and strategic languages such as Chinese and Arabic as well as minor languages are in the early stages of development.

Two major projects in the U.S. and Europe have joined in a collaboration to work toward achieving interoperability among language resources. In the U.S., a project entitled ”Sustainable Interoperability for Language Technology” (SILT) has been funded by the National Science Foundation under the INTEROP program, and in Europe, FLaReNet Fostering Language Resources Network has been funded by the European Commission under the eContentPlus framework. This international collaborative effort involves members of the language processing community and others working in related areas to build consensus regarding the sharing of data and technologies for language resources and applications, to work towards interoperability of existing data, and, where possible, to promote standards for annotation and resource building. In addition to broad-based US and European participation, we are seeking the participation of colleagues in Asia.

To ensure full community involvement and consolidation of effort, SILT and FLaReNet are establishing ties with major ongoing projects and consortia, including the International Standards Organization TC37 SC4 (Language Resource Management)[1], The World Wide Web Consortium (W3C), the Text Encoding Initiative, the ACL Special Interest Group on Annotation (SIGANN)[2], and others. The ultimate goal is to create an Open Language Infrastructure (OLI) that will provide free and open access to resources, tools, and other information that support work in the field, in order to facilitate collaboration, accessibility for all members of the community, and convergence toward interoperability.

The following sections outline the goals of SILT and FLaReNet.

2 SILT

The creation and use of language resources spans several related but relatively isolated disciplines, including NLP, information retrieval, machine translation, speech, and the semantic web. SILT's goal is to turn existing, fragmented technology and resources developed within these groups in relative isolation into accessible, stable, and interoperable resources that can be readily reused across several fields.

The major activities of the effort are:

- carefully surveying the field to identify the resources, tools, and frameworks in order to examine what exists and what needs to be developed, and to identify those areas for which interoperability would have the broadest impact in advancing research and development and significant applications dependent on them;

- identifying the major efforts on standards development and interoperable system design together with existing and developing technologies, and examining ways to leverage their results to define an interoperablity infrastructure for both tools and data;

- analyzing innovative methods and techniques for the creation and maintenance of language resources in order to reduce the high costs, increase productivity, and enable rapid development of resources for languages that currently lack them;

- implementing proposed annotation standards and best practices in corpora currently under development (e.g., American National Corpus[3], TimeBank[4]) to evaluate their viability and feed into the process of further standards development, testing, and use of interoperability frameworks (e.g., GATE[5], UIMA[6]) and implementation of processing modules, and distributing all software, data, and annotations.

- ensuring the broadest possible community engagement in the development of consensus and agreement on strategies, priorities, and

[1] http://www.tc37sc4.org

[2] http://www.cs.vassar.edu/sigann

[3] http://www.anc.org

[4] http://www.timeml.org/site/timebank/timebank.html

[5] http://gate.ac.uk

[6] http://www.oasis-open.org/committees/uima/

best approaches for achieving broad interoperability by means of sessions, open meetings, and special workshops at major conferences in the field, together with active maintenance of and involvement in open web forums and Wikis;

- providing the technical expertise necessary to turn consensus and agreement into robust interoperability frameworks along with the appropriate tools and resources for their broad use and implementation by means of tutorials and training workshops, especially for undergraduate and graduate students in the field.

3 FLaReNet

The multilingual Europe urgently needs language technologies in order to bridge its language barriers. In order to achieve better quality and fast development of language technologies that seamlessly work on all devices, for spoken and written language alike, the European scenario now needs a coherent and unified effort. The demand for cross-lingual technologies is pressing, the expectations are high, and at the same time, the field is suffering from fragmentation, lack of vision and direction. The main objective of FLaReNet is to steer the process that in the near future will define the actors, the overall direction and the practical forms of collaboration in language technologies and their ”raw material”, language resources. Under this respect, the goals of FLaReNet lie at a higher level than those of SILT, as they are oriented towards consolidating a community around a number of key topics that, in the end, will allow networking of language technology professionals and their clients, as well as easy sharing of data, corpora, language resources and tools.

From this perspective, FLaReNet has three main lines of action:

The creation and mobilization of a unified and committed community in the field of Language Resources and Technologies. To this end, FLaReNet is bringing together leading experts of research institutions, academies, companies, funding agencies, public and private bodies, both at European and international level, with the specific purpose of creating consensus around short, medium and long-term strategic objectives. The Network is currently composed of around 200 individuals belonging to academia, research institutes, industries and government.

The identification of a set of priority themes on which to stimulate action, under the form of a roadmap for Language Resources and Technologies. In order to avoid scattered or conflicting efforts, the major players in the field of Language Resources and Technologies need to consensually work together and indicate a clear direction of action and a shared policy for the next years. This will take the form of identification of priorities of intervention as well as short, medium, and long-term strategic objectives at all levels, from research directions to implementation choices, from distribution and access policies to the landscape of languages, domain and modalities covered by Language Resources and Technologies.

The elaboration of a blueprint of priority areas for actions in the field and a coherent set of recommendations for the policy-makers (funding agencies especially), the business community and the public at large. Whatever action cannot be implemented on a long term without the help of the necessary financial and political framework to sustain them. This is even most true for actions regarding Language Resources that typically imply a sustained effort at national level. To this end, the FLaReNet Network will propose the priority themes under the form of consensual recommendations and a plan of action for EC Member States, other European-wide decision makers, companies, as well as non-EU and International organizations.

The following Thematic Areas are currently covered by FLaReNet:

- The Chart for the area of LRs and LT in its different dimensions
- Methods and models for LR building, reuse, interlinking, and maintenance
- Harmonisation of formats and standards
- Definition of evaluation and validation protocols and procedures
- Methods for the automatic construction and processing of Language Resources

FLaReNet builds upon years of research and development in the field of standards and language resources, as well as on the achievements (both in terms of results and community awareness), of past EU projects such as EAGLES[7], ISLE[8],

[7] http://www.ilc.cnr.it/EAGLES/home.html

[8] http://www.ilc.cnr.it/EAGLES/isle/ISLE_Home_Page.htm

INTERA[9], and LIRICS[10]. Close collaboration is also established with many relevant ongoing EU projects, such as CLARIN[11].

[9] http://www.elda.org/intera

[10] http://lirics.loria.fr/

[11] http://www.clarin.eu