Harmonization of Formats and Standards

Trends, Priorities and Strategies

Nancy Ide
Vassar College, USA

Focus on Interoperability

- Interoperability of resources, tools, and frameworks now recognized as the most pressing need for language processing research
- Motivations:
  - Need to create and merge annotations at different linguistic levels in order to study interactions and interleave processing
  - Need to develop data and tools for emerging and strategic languages such as Chinese and Arabic, and minor languages
  - Need to make a major leap in the productivity of NLP research and language processing capabilities

Activities

- **Formation of ISO TC37 SC4** to develop a linguistic annotation framework and standard representation formats for various types of linguistic annotation
- Global efforts to create **linked wordnets and framenets**
- Development and harmonization of **systems and frameworks for linguistic annotation** (e.g., GATE, Callisto, UIMA)
- **Recent major meetings** devoted to resource interoperability
  - E-MELD, TILR
  - International conference devoted to language resource interoperability (ICGL)
  - Multiple workshops at major conferences addressing issues of standards for representation formats and linguistic categories

Activities

- Establishment of **registries and catalogues for linguistic categories** (e.g., ISO TC37 SC4 data category registry) and annotation schema (e.g., UIMA component registry)
- **U.S.-funded efforts** to merge and/or harmonize linguistic annotations at different levels (OntoNotes, Unified Linguistic Annotation), and different phenomena (WordNet and FrameNet)
- **EU-funded effort** to create a common resource and infrastructure for the humanities and social sciences (CLARIN)
- **Formation of an ACL special interest group** (SIGANN), with a primary aim to work toward the development of standards for representing and designating linguistic information
- Independent work within the Semantic Web community on interoperability of ontologies

Where We Are

- These efforts developed in (some) isolation
- But now:
  - Enough convergence of opinion and practice that the steps required to bring the pieces together are beginning to emerge
  - Advances in technology (e.g., Semantic Web technologies) and the emergence of distributed computing have opened up possibilities for interlinkage of data, annotations, lexicons, ontologies, etc.

Where We Are

- The time and circumstances are ripe to move toward establishing and implementing standards and technologies necessary to ensure language resource interoperability in the future

- Can only be achieved through a **coordinated, community-wide effort** to ensure comprehensive coverage and widespread acceptance

INTEROP

- In parallel with submission of proposal for FLaReNet to EU, submission of proposal in the US to National Science Foundation's INTEROP program
- PIs: Nancy Ide, James Pustejovsky
- Letters of commitment to involvement from wide range of US researchers
- Funded for 3 years, start sometime before August 1

INTEROP Goals

- Survey of resources, tools, and frameworks
  - Examine what exists and what needs to be developed
  - Identify areas for which interoperability would have the broadest impact in advancing R&D
- Identify major standards/interoperability efforts and existing and developing technologies
  - Examine ways to leverage results to define an interoperablity infrastructure for tools and data
- Analyze innovative methods and techniques for the creation and maintenance of language resources in order to
  - Reduce high costs
  - Increase productivity
  - Enable rapid development of resources for new languages

INTEROP Goals

- Implement proposed standards and best practices in corpora currently under development (e.g., American National Corpus, TimeBank)
  - Evaluate their viability
  - Feed into the process of standards development
  - Test and use interoperability frameworks (e.g. UIMA), and implement processing modules
  - Distribute all software, data, and annotations

Short Term Benefits

- Ability to combine annotations of different phenomena produced by different groups in order to study interactions among linguistic levels
- Creation of lexical/semantic/ontological resources that include information relevant to different sub-domains (speech, machine translation, information retrieval)
- Provide substantially increased access to resources and tools for members of the entire community
- Enable rapid development of resources for new languages
- Enhance the teaching of NLP

Long Term Vision

- Creation of a web-based “resource grid”
  - Lexical, semantic, and ontological resources are interlinked – can serve as the reference for linguistic annotations
- Eventual creation of
  - Massive, distributed, interlinked network of linguistic data and information
  - Web services to accomplish linguistic processing “on the fly” in real time

INTEROP Goals

- Ensure broad community engagement in the development of consensus
  - Hold sessions, open meetings, and special workshops at major conferences
  - Actively maintain and be involved in open web forums and wikis
- Provide technical expertise to turn consensus and agreement into robust interoperability frameworks, and tools and resources for their use
  - Hold tutorials and training workshops, especially for students in the field

Some Aspects

- Harmonization of formats for linguistic data and annotations
- Harmonization of descriptors in linguistic annotation

Formats: The Past 20 Years

1987	TEI
1994	MULTEXT, CES
~1996	XML
2000	ISO TC37 SC4
2001	LAF model introduced
now	LAF/GrAF, ISO standards

Myriad of formats

Myriad of formats

Actually…

- Things are better now
  - XML use
  - Moves toward common models, especially in Europe
  - US community seeing the need for interoperability
  - Emergence of common processing platforms (GATE, UIMA) with underlying common models

Reality

- Harmonization of formats moving in the right directions, but will take some time
- Use of standards always driven by processing capabilities
  - We may think globally, but we act locally
- Harmonization of formats foreseeable, but harmonization of descriptors is more difficult

What Next?

- We now have the will to work toward interoperability
- Unlikely we will see real harmonization after the efforts of the next few years
- But we need to make steps
  - E.g. immediate: develop evaluation criteria / requirements for LRs (“ISO 9000 approved”)
  - Do the thinking and development / testing provided for in projects like FLaReNet and INTEROP

Recommendations

- Need a paradigm shift in our thinking
  - Going along thinking in the same terms probably won't get us far
  - A need to get another (more) perspective(s)
- The effort needs to be much broader in scope
  - Need to involve other communities
    - Semantic web and W3C, several fields in computer science…
- The danger:
  - Some areas are mature enough for standardization, but trying to standardize otherscould backfire