Annotation Science

From Theory to Practice and Use

Nancy Ide
Department of Computer Science
Vassar College
Poughkeepsie, New York 12604 USA
ide@cs.vassar.edu

Introduction

Linguistically-annotated corpora are an increasingly critical resource for research in linguistics and computational linguistics. As a result, there is now unprecedented interest in the development of annotated language resources as well as establishing standard practices for corpus annotation. The ultimate goal is to enable interoperability among annotations for different linguistic phenomena for the same language, together with linguistic annotations applied to different languages and modalities. Over the past twenty years, numerous efforts have contributed to the development of standards for representing and processing language resources, during which time we have seen substantial changes in the ways in which language resources are used and accessed and a corresponding development of the supporting technologies. We are now at a point where the collective experience of the past two decades, together with a clear idea of what the current technology both enables and demands of language processing research, puts us in a position to take a major step toward the interoperability of resources, including those involving multiple modalities and languages.

A bit of history

We can trace a history of "big ideas" in the area of linguistic resource creation and use. The first may be the TEI, which intended to standardize the representation of annotations (by which I mean any descriptive information added to the data), relying on, first, SGML and then its successor, XML. The Corpus Encoding Standard (CES) and its XML instantiation (XCES) applied and extended the TEI to provide a single representation format for linguistic annotations. It is important to note that the CES addressed the difficult problem of drawing the line between prescribing what might be called "annotation content categories" differently from the TEI; the TEI provides multiple options for annotating a given phenomenon (e.g., vs. <p>), whereas the CES provided a single option. However, for content categories describing linguistic phenomena—such as a morpho-syntactic description or a syntactic label—the CES backed off from the prescriptive approach, providing only generic category tags such as <msd> and recommending that the specific linguistic annotation category be given in an attribute or as tag content. Specifications for specific linguistic category descriptions were left to projects such as EAGLES/ISLE, of which the CES was a part.

The CES also introduced another "big idea" for representing linguistic annotations: the notion of "remote markup" or, as it is more commonly referred to now, "standoff annotation". This idea took hold immediately within the resource-building community

and is accepted as standard practice today. The use of standoff annotation meant that users did not modify the original data, but instead linked annotation information stored in other documents to it. It also made it easier to represent complex annotations, since annotations can be treated as data (e.g., by appearing as tag content in an XML structure), and avoided most problems with overlapping hierarchies among annotations of the same or different types, since each annotation could be in a separate document. Bird and Liberman's Annotation Graphs (Bird and Liberman, 2001) provided a formal model for standoff annotations, but suffered from the drawback that annotations could not be linked to other annotations, thus making the representation of hierarchical annotations (e.g., syntactic constituency) problematic.

Another "big idea" that did not originate in the computational linguistics community, but which significantly impacted our thinking about representing annotations, is the Semantic Web and its supporting technologies, RDF/RDFS and OWL. The notion of linking pieces of information that may be distributed across the web provided a different model for linguistic resources, in which annotations of different documents may reference the same "object" providing the annotation content and/or point to lexicons or knowledge bases that may themselves be linked to, say, multi-lingual resources. Similarly, the development of OWL provided new motivation for the development of common ontologies and provided a means by which multiple resources could be associated with them, leading, in part, to work on ontologies for linguistic content categories. These ideas were not new *per se*, but the availability of the technology to enable them—not to mention the web itself—caused them to take on a more central role in the ways in which we create and use annotated language resources. The existence of the web and web technologies also motivated the recent interest in global linkage of multi-lingual and multi-modal resources, and, perhaps most importantly, enabled instantaneous sharing of language resources and software developed in other parts of the world. The need to use and integrate such resources served to make even the most confirmed skeptics acknowledge the need for standardization.

Annotation Science Today

Far from the situation 20 years ago, when annotations were added to data without much thought about their physical format or the repercussions of the choice of content categories, a "science" of annotation has now evolved that reflects the collective experience within the community. This new science includes the study and development of precise criteria for corpus design, appropriate statistics for measuring inter-annotator agreement and confidence, and means to define a set of annotation categories that reflect an underlying linguistic theory. It is also concerned with the design of an architecture for annotated resources that supports interoperability, and its implementation in systems and frameworks that support the creation and exploration of annotations.

Work in these areas, together with several *de facto* approaches that have evolved from the "big ideas" discussed earlier, has led to a much clearer picture of potential standard ways to create, represent, and manipulate linguistically annotated data. Recognizing this, the International Standards Organization (ISO) recently formed a sub-committee on Language Resource Management (ISO TC37 SC4) to define standards for representing linguistic annotations and other resources, by incorporating *de facto* standards and "best

practices" into a coherent whole. The core of this work is the definition of a Linguistic Annotation Framework (LAF) that is defined broadly enough to accommodate all types of linguistic annotations and provides means to represent precise and potentially complex linguistic information.

The Linguistic Annotation Framework

LAF development has been guided by a few general requirements. First, it is essential that LAF accommodate all varieties of annotation and data (including, e.g., time-stamped speech, streamed data, multi-lingual and multi-modal data, etc.). In addition, LAF must enable users to represent their data and annotations in a variety of formats of their own choosing. Finally, it must be easy to use so that the community will adopt it.

The definition of a standard seems at odds with the requirement that users can use any format they choose. However, there is a *quid pro quo*: user-defined formats must conform to a feature structure-based abstract data model defined by LAF. The abstract model is instantiated by a "dump" format that is intended to function in the same way as an interlingua functions for machine translation--i.e., as a representation of universal concepts into and out of which realizations in different languages are mapped for the purposes of translation. Thus, users may use XML or any other format such as LISP-like structures or tab-delimitation, as long as the information in the user's annotation format is automatically mappable to the abstract model.

The abstract model is based on the principle that the structure and content (i.e., the linguistic information associated with the primary data) of annotations are separated. While this may seem to be a simple idea, in many annotation schemes content and structure are not clearly differentiated and, as a result, structural relations among parts of the annotation content are ambiguous. The most obvious example is LISP-like formats, which use parentheses to group information, with no indication of whether the group represents constituency, an inclusive list, a prioritized list, a set of alternatives, etc. The only way to determine which applies is to examine the data; if, for instance, the list describes syntactic frames or part of speech for a given lexical item, it is probably a set of alternatives, but human knowledge is required to decide this and program a script to treat it appropriately. LAF requires that all annotation information included in the original format be made explicit in the dump format representation; this way, fully automatic transduction from the dump format representation or other user formats is ensured.

In principle, users will never deal directly with, or even see, the dump format, and therefore the primary considerations for its design are to maximize processing efficiency and consistency, ensure that processing is unambiguous, and ensure that the mapping from user formats is not overly complex. The dump format represents an annotation as a directed graph referencing *n*-dimensional regions of primary data as well as other annotations. In the primary data, the nodes of the graph are virtual, located between each "character" in the primary data, where a character is defined to be a contiguous byte sequence of a specified length.[1] When an annotation references another annotation document rather than primary data, the nodes are the edges within that document that have been defined over the primary data or other annotation documents. That is, given a

[1] For text, the default is UTF-16.

graph, *G*, over primary data, we create an *edge graph G'* whose nodes can themselves be annotated, thereby allowing for edges between the edges of the original graph *G*. Edges are labeled with feature structures containing the annotation content relevant to the data identified by the edge.

The dump format is instantiated in XML. ISO TC37 SC4 has collaborated with the Text Encoding Initiative (TEI) Consortium to adapt and revise the TEI's specifications for representing feature structures in XML[2]. The ISO/TEI specifications implement the full power of feature structures and define inheritance, unification, and subsumption mechanisms over the structures, thus enabling the representation of linguistic information at any level of complexity. The specifications also provide a concise format for representing simple feature-value pairs, which suffices to represent many annotations.

It is important to note that in principle, the dump format places no restrictions on annotation content (i.e., the categories and values in an annotation); annotation content is effectively user-defined, taken directly from the user's original annotation. However, it is obvious that harmonization of content categories is a critical next step toward standardizing annotations. LAF is addressing this far more controversial and problematic issue separately. Two major activities within ISO TC37 SC4 are aimed at harmonization of annotation content: (1) definition of user annotation formats for different annotation levels[3], and (2) creation of a Data Category Registry (DCR) containing pre-defined data elements and schemas that can be used directly in annotations (Ide and Romary, 2004). The DCR includes atomic data categories (both category names and values) that may be referenced directly in user annotations, or to which a mapping from user–defined categories can be included in the dump format document. In addition, feature structure libraries that can be referenced directly in both user and dump format annotations are under development.

Conclusion

The Linguistic Annotation Format brings together the "big ideas" that have influenced our approach to the annotation task over the past 15-20 years. So far, it seems capable of enabling the interoperability among language resources that is increasingly crucial to the development of natural language processing applications. As a test, LAF is being used to represent the American National Corpus[4], which includes a broad variety of annotation types. It is also isomorphic to representation schemes used in widely-used annotation systems such as UIMA[5]. At the least, LAF represents a major step towards resource interoperability that has led to unprecedented collaboration among annotators and system developers throughout the world.

---

[2] See ISO TC37 SC4 document N188, Feature Structures-Part 1: Feature Structure Representation (2005-10-01), available at http://www.tc37sc4.org/

[3] Draft documents and working papers for the various areas, including morpho-syntactic annotation (ISO TC37 SC4 document N225), syntactic annotation (ISO TC37 SC4 document N244), word segmentation (ISO TC37 SC4 document 233), etc. are available at http://www.tc37sc4.org/.

[4] http://AmericanNationalCorpus.org

[5] http://www.research.ibm.com/UIMA/

References

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication,* 33:1-2, 23-60.

Ide, N., Romary, L. (2004). A Registry of Standard Data Categories for Linguistic Annotation. In *Proceedings of the Fourth International Language Resources and Evaluation Conference* (LREC), Lisbon, pp. 135-39.