An Efficient Approach to Gold-Standard Annotation: Decision Points for Complex Tasks

Julie Medero, Kazuaki Maeda, Stephanie Strassel, Christopher Walker

Linguistic Data Consortium
University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA USA
{jmedero, maeda, strassel, chwalker}@ldc.upenn.edu

Abstract

Inter-annotator consistency is a concern for any corpus building effort relying on human annotation. Adjudication is as effective way to locate and correct discrepancies of various kinds. It can also be both difficult and time-consuming. This paper introduces Linguistic Data Consortium (LDC)'s model for decision point-based annotation and adjudication, and describes the annotation tools developed to enable this approach for the Automatic Content Extraction (ACE) Program. Using a customized user interface incorporating decision points, we improved adjudication efficiency over 2004 annotation rates, despite increased annotation task complexity. We examine the factors that lead to more efficient, less demanding adjudication. We further discuss how a decision point model might be applied to annotation tools designed for a wide range of annotation tasks. Finally, we consider issues of annotation tool customization versus development time in the context of a decision point model.

1. Introduction

Consistency issues in human annotation tasks are frequently addressed by creating gold standard data that has been labeled for the same task by multiple annotators, and then adjudicated by a senior annotator to resolve discrepancies. Adjudication is an effective way to locate and correct both detection (finding an instance of a targeted linguistic feature) and characterization (describing how that feature is realized) mismatches. The resulting gold standard data can be used to benchmark system performance and human consistency. Moreover, the adjudication process itself provides insight into the nature of annotation discrepancies, and can suggest improvements to annotation guidelines and annotator training.

Because the cost of adjudicating large volumes of data may be prohibitive, it is frequently only applied to a small subset of all annotated data. For instance, when creating linguistic resources for common task technology evaluations, Linguistic Data Consortium (LDC) typically produces gold standard files for all text data, but only 10-15% of training data. In 2005, however, the Automatic Content Extraction (ACE) Program required that all training and evaluation data – over a million words in all – be dually annotated and adjudicated. This requirement demanded a new approach to adjudication that was both faster and easier for the annotator. A new adjudication tool designed using decision-points made it possible to provide large quantities of adjudicated data

2. Adjudication and the ACE Task

2.1. The ACE Annotation Task

In ACE, annotators detect, characterize and coreference entities, relations and events, making numerous judgments about the linguistic and semantic properties of each item. Annotations are structurally complex and interdependent. A single mention of an event includes references to each of the entities that participate in the event. Depending on the event type, different restrictions exist on the types of the entities that can fill each role in the event's argument structure. For example, in a Life.Die event, one of the allowable arguments is a Victim-Arg (LDC, 2005). This argument slot can only be filled by a Person entity. If an entity is incorrectly tagged as an Organization instead of as a Person, it will not be an allowable option for the Victim-Arg slot of a Life.Die event. Decisions about the type of an entity, then, also affect every event that the entity participates in.

2.2. Adjudication for ACE

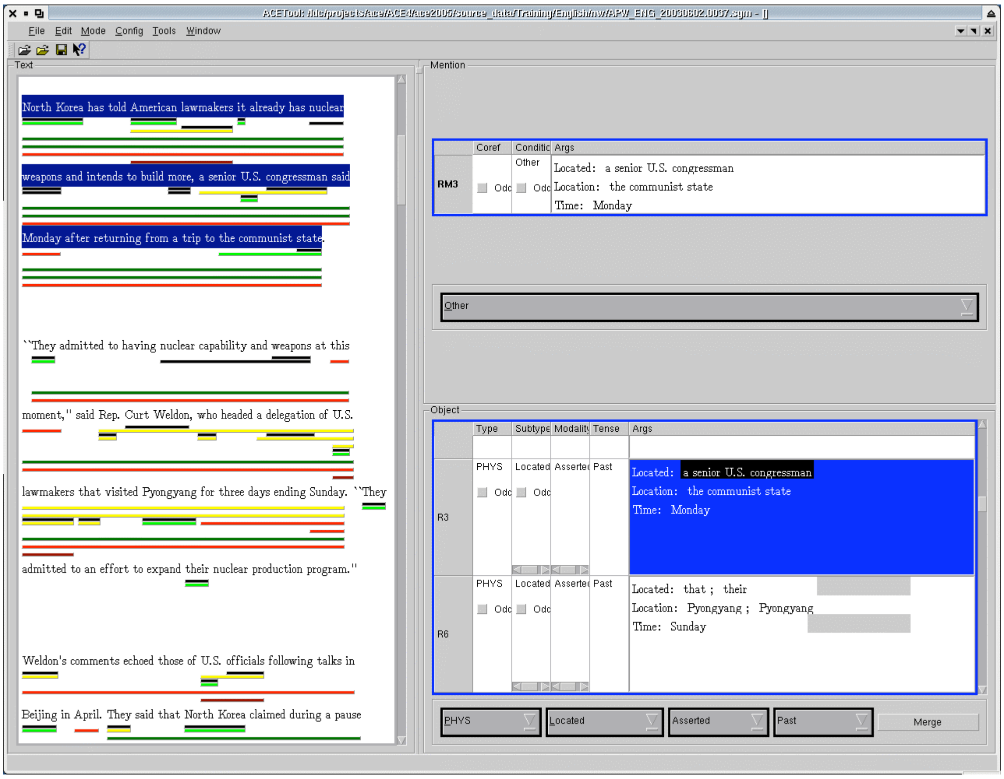
Though some adjudication tasks (for instance, comparing words in two versions of a transcript) can be straightforward, it is much more difficult when applied to a multi-faceted annotation task like ACE. Adjudicating event mentions, for example, involves consideration of not only the attributes of an event, but of all the entities that participate in the event. When approached as a simple list of disagreements, the ACE adjudication task presents a substantial cognitive burden to the human annotator, who must keep all of these attributes and inter-dependencies in mind while resolving each discrepancy. Furthermore, resolution of one discrepancy may affect the analysis of other discrepancies, so the annotator must always be alert to the trickle-down effect of each judgment on the rest of the data. For instance, changing the type of an entity to make it a valid participant in one event might have the unintended consequence of making it an invalid option for a participant in another unrelated event.

With this approach, adjudication is time-consuming and fatiguing. Annotators report both frustration with the process and lack of confidence in the results it produces. As a result, adjudication was not included in the production pipeline for ACE Training data in 2004, and only a small fraction of the Evaluation files for that year were adjudicated.

3. Decision Points

3.1. What are Decision Points?

Decision points refer to the series of judgments that a human must make in order to complete an annotation task.


ACETool: /ldc/projects/ace/ACE4/ace2005/source_data/Training/English/nw/APW_ENG_20030602.0037.sgm - []
File
Edit
Mode
Config
Tools
Window
Text
Mention
Object
Other
PHYS
Located
Asserted
Past
Merge

North Korea has told American lawmakers it already has nuclear weapons and intends to build more, a senior U.S. congressman said Monday after returning from a trip to the communist state.

``They admitted to having nuclear capability and weapons at this moment,'' said Rep. Curt Weldon, who headed a delegation of U.S. lawmakers that visited Pyongyang for three days ending Sunday. ``They admitted to an effort to expand their nuclear production program.''

Weldon's comments echoed those of U.S. officials following talks in Beijing in April. They said that North Korea claimed during a pause

	Coref	Conditic	Args
RM3	□ Odc	Other □ Odc	Located: a senior U.S. congressman Location: the communist state Time: Monday

	Type	Subtype	Modality	Tense	Args
R3	PHYS □ Odc	Located □ Odc	Asserted	Past	Located: a senior U.S. congressman Location: the communist state Time: Monday
R6	PHYS □ Odc	Located □ Odc	Asserted	Past	Located: that ; their Location: Pyongyang ; Pyongyang Time: Sunday

Figure 1: ACE Annotation Tool

To annotate an ACE entity, for instance, annotators must answer a series of questions: Is there an annotation object here? What is its extent? What is its head? Is it a name, nominal, or pronoun? And so on. Each of these questions constitutes a decision point; if one of the decisions is left unmade, the annotation is incomplete. In some cases, the order of the decision points is crucial to the annotation process; for instance, the set of valid subtypes for an entity can not be determined until after the type has been set.

Introducing the notion of decision points can significantly reduce cognitive load on human annotators. While standard adjudication requires annotators to identify and resolve discrepancies across varying features of interdependent annotations, enforcing decision points allows the annotator to focus on one kind of feature at a time.

3.2. Decision Points in Annotation

Decision points are already utilized in the basic ACE task, where the ACE Toolkit (Figure 1) steps annotators through each stage of annotation, prompting them for a response at each decision point. To annotate an entity, for instance, an annotator must select the extent of the entity mention. They are then prompted for the extent of the mention's head. Next, they are asked to select a mention type (name, nominal or pronoun). This process is continued until all of the decisions relevant to the annotation have been made.

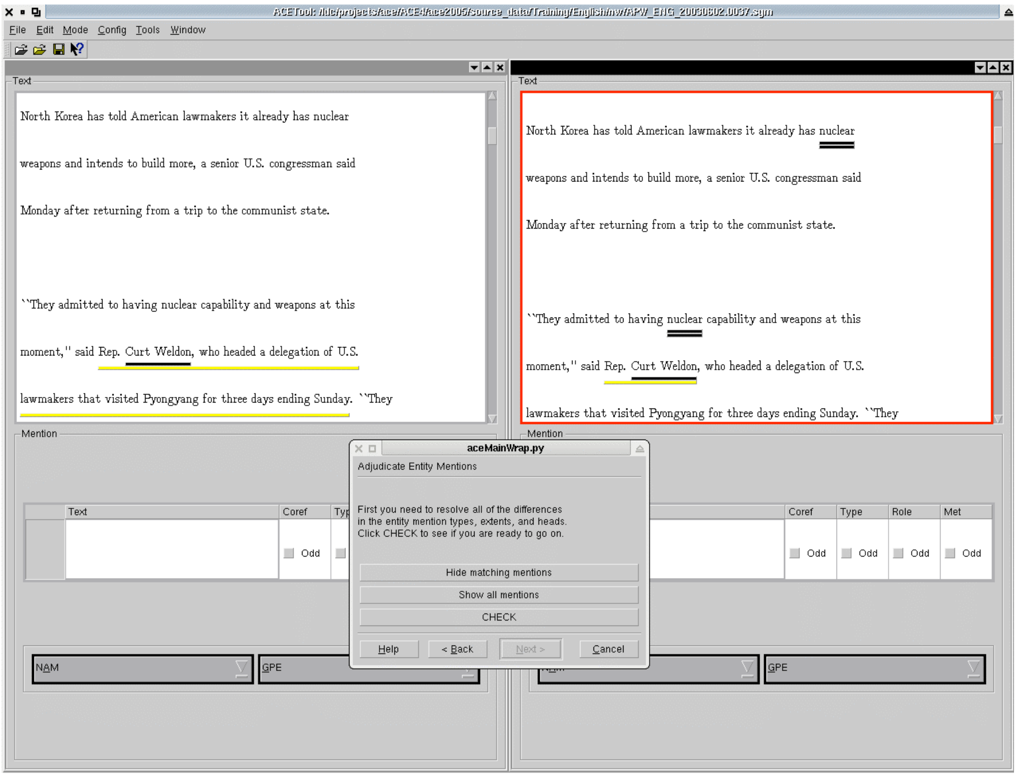
In this way, the ACE Toolkit prevents the annotator from making decisions out of their logical order; a valid annotation must be supplied at each stage, and those decisions determine in part what options are available at future decision points. The resulting annotations have fewer errors: no decisions are left unmade (for example, an entity without a type), and dependencies between decisions are preserved (so that it is impossible, for example, to create an entity whose type and subtype are logically disallowed).

3.3. Decision Points in Adjudication

This same approach can help to constrain and enhance the adjudication process by imposing logical structure on the annotator's decisions. In the ACE Toolkit, a wizard imposes this structure. The adjudication wizard takes each decision point for a task (for instance, all of the entity type decisions) and presents the human adjudicator with just the set of discrepancies that are a result of decisions made for that point. The annotator is required to resolve each discrepancy associated with that decision point before moving on to the next.

A benefit of this approach is that the adjudication wizard is able to offer an increasingly narrow range of options as the adjudication process progresses. The first stage of adjudication asks the human adjudicator to resolve entity mention extents. At this stage, the wizard can hide mentions that already match, but the human adjudicator must make all other mention-pairing decisions. Figure 2 shows the ACE Toolkit with the adjudication wizard active during the mention extent-resolution process. While the tool is not yet able to fully map annotations to characterize discrepancies for the human adjudicator, it is able to substantially limit the amount of information displayed on the screen at one time, as demonstrated by the decrease in underlined areas of the text in the tool.

Once mention extents and coreference have been resolved, on the other hand, the wizard is able to very accurately identify remaining discrepancies. For example, Figure 3 shows the toolkit with the adjudication wizard prompting for resolution of the type and subtype of an entity whose mentions have already been resolved.



ACETool: /ldc/projects/ace/ACE4/ace2005/source_data/Training/English/nw/APW_ENG_20030602.0037.sgm

North Korea has told American lawmakers it already has nuclear

weapons and intends to build more, a senior U.S. congressman said

Monday after returning from a trip to the communist state.

``They admitted to having nuclear capability and weapons at this

moment,'' said Rep. Curt Weldon, who headed a delegation of U.S.

lawmakers that visited Pyongyang for three days ending Sunday. ``They

North Korea has told American lawmakers it already has nuclear

weapons and intends to build more, a senior U.S. congressman said

Monday after returning from a trip to the communist state.

``They admitted to having nuclear capability and weapons at this

moment,'' said Rep. Curt Weldon, who headed a delegation of U.S.

lawmakers that visited Pyongyang for three days ending Sunday. ``They

aceMainWrap.py

Adjudicate Entity Mentions

First you need to resolve all of the differences
in the entity mention types, extents, and heads.
Click CHECK to see if you are ready to go on.

Figure 2: ACE Adjudication Tool

Under traditional adjudication strategies, if two single-mention entities differ in more than one way (for example, in mention type and entity subtype), they are presented as a single discrepancy. With decision point-based adjudication, each differing feature is presented as a separate, simple discrepancy-resolving decision. This approach decreases the cognitive load on the annotator in two ways: it minimizes the number of attributes and dependencies that the annotator needs to mentally track at each step; and it ensures that earlier decisions are correct before addressing later decisions.

4. Discussion

4.1. Impact on Efficiency

The reduction in cognitive load for the annotator is coupled with a great reduction in time and frustration. Decision point-based methods have allowed ACE annotators to complete adjudication even more quickly than basic annotation. Figure 3 shows the relative times needed for adjudicating newswire and broadcast news files for the English, Chinese and Arabic ACE Evaluation[1] datasets in 2004[2] and 2005. Adjudication was faster than annotation for all languages and genres in 2005 (that is, adjudication rate / annotation rate < 1). The only dataset that did not see an improvement in relative annotation rate from 2004 to 2005 was the English NW data. This can be explained by the already fast adjudication rates for the 2004 English data, which was completed exclusively by two full-time, senior annotators. A full staff of part-time annotators adjudicated the 2005 English data.

The improved adjudication rate from 2004 to 2005 is particularly noteworthy given the complexities that were added to the task. In 2004, files were only annotated for Entities and Relations. In 2005, they were annotated for Entities, Values, Relations and Events. As discussed previously, errors in entity annotation are propagated to event annotation as well, so the overall complexity of the adjudication task increased substantially. In addition, documents that were annotated in 2005 were subject to a rigorous data selection process to ensure a high density of annotatable objects. As a result, the actual number of annotations compared during the adjudication process in 2005 was substantially larger than the number of annotations compared during the same process in 2004. Despite the added task complexity and density, though, designing a tool with decision points made it possible for a larger group of annotators to efficiently complete the adjudication task.

4.2. Development Time

In any annotation project, a balance must be reached between tool development time and tool customization. Highly flexible and configurable annotation tools can minimize the upstart time needed to reach a point where production-level annotation can begin for a project. A tool highly customized to the thought process and extensive dependencies of a specific task, like the ACE Toolkit, on the other hand, requires substantially more programmer time to implement.

[1] In 2005, LDC did not produce data for the ACE Arabic Evaluation dataset. Consequently, numbers reported here are for ACE Training files annotated during the same period that English and Chinese Evaluation files were being produced.

[2] Due to time constraints and annotator dificulties with the tool, no Arabic files were adjudicated in 2004, so only 2005 numbers are reported here for Arabic. In 2005, we were able to successfully adjudicate over 100,000 words of Arabic data.

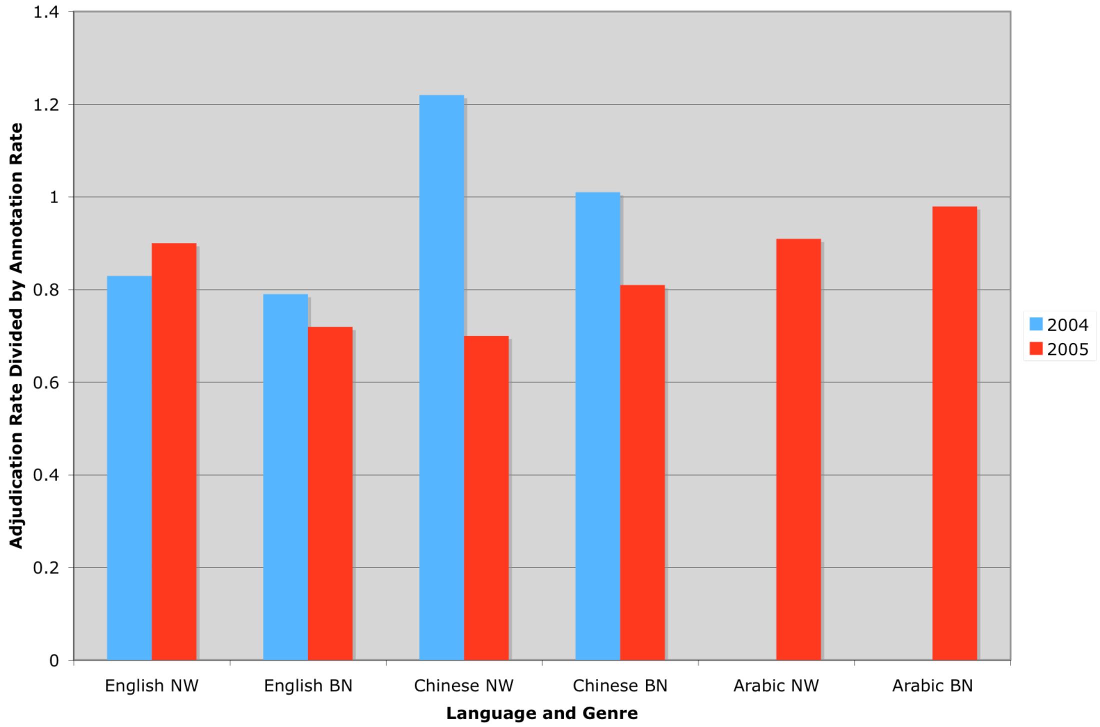


Figure 3: ACE Adjudication Rates Relative to Annotation Rates

Given the large quantity of data that was annotated for the 2005 ACE Program, increased development time was more than made up for by the increased efficiency of the annotation itself. At the same time, though, we are always looking for ways to minimize tool development time. Like other tools developed at LDC, the ACE Toolkit makes use of the Annotation Graph Toolkit (AGTK) (Maeda et. Al, 2006; Maeda, Strassel 2004). By building off of existing GUI components and the AG Library's API, we are able to minimize the programmer effort that goes into developing the user interface and storing the annotation information. This allows us to focus our development effort on enforcing decision points and making the annotation process as straightforward as possible for human annotators.

The ideal, of course, would be to have a highly configurable annotation tool that, once configured, would understand and enforce the decision points of an annotation task. Such a tool would allow for rapid deployment of tools for new projects without losing the efficiency benefits of a highly customized tool. We have made some progress in this direction with the ACE Toolkit. The allowable options for each decision point are now stored in an XML configuration file that makes it easy to make slight adjustments to the task to accommodate different task specifications as well as differences in the task specification between languages. Exploration of a way to represent the decision points themselves in a configuration file so that decision orders and dependencies could be easily customized is an area that warrants further study.

5. Conclusion

Developing highly configurable, customized annotation software facilitates efficient creation of linguistic data while serving as an aid to annotation consistency. Annotation and adjudication tools are most beneficial when they model the thought process an annotator goes through when deciding how to label a piece of linguistic data. Relying on the notion of decision points in the ACE annotation and adjudication tasks has improved annotation efficiency and quality, and resulted in far less annotator fatigue and frustration than more standard approaches. Given the success of this approach for ACE we plan to extend it to other cases of complex inter-dependent annotation tasks.

6. References

Linguistic Data Consortium, 2005. English-Events-Guidelines_v5.4.3. http://projects.ldc.upenn.edu/ace/docs/English-Events-Guidelines_v5.4.3.pdf

Maeda, Kazuaki, Haejoong Lee, Julie Medero and Stephanie Strassel, 2006. A New Phase in Annotation Tool Development at the Linguistic Data Consortium: The Evolution of the Annotation Graph Toolkit. In *Proceedings of the Fifth International Conference on Language Resources and Evaluations.*

Maeda, Kazuaki and Stephanie Strassel, 2004. Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.