Introduction to Manual Annotation

This document introduces the concept of annotations, their uses and the common types of manual annotation projects.

This is a supplement to project-specific guidelines that describe the goals and instructions for a particular annotation project. Project guidelines will provide detailed instructions for annotators and/or curators, including examples of the desired annotations. Annotators, curators, and project managers should also consult the GATE Teamware User Guide *for information on how to annotate with that tool.*

April 2010

Matthew Petrillo

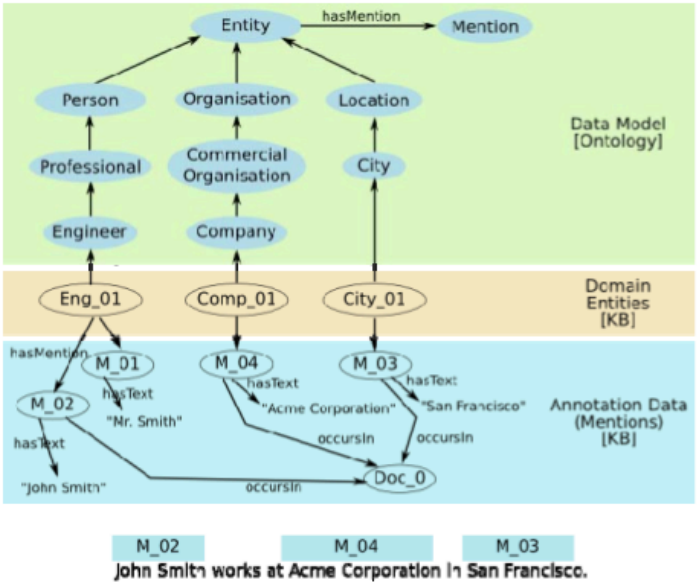
Jessica Baycroft

Introduction to Annotations

Annotation is a methodology for adding information to a document at some level—a word or phrase, paragraph or section or the entire document. This information is "meta-data," that is, data about other data.

The difference between annotation and other forms of meta-data is that an annotation is grounded to a specific point in a document. For example, one might consider a folder name on a computer as meta data for the files in that folder. So, a folder labeled "holiday 2008" might hold files of photographs taken on holiday. The folder name is a form of meta-data. But, when an image file is taken out of the folder, it becomes separated from that meta data and thus loses some valuable context. Many desktop image managers compensate by including meta-data tags that can be added to the image file. Thus, the data stays with the image, but you need a specialized picture viewer to see that meta-data and use it to sort, categorize or find pictures. Some programs allow tags to be applied to specific regions of an image. Facebook has such an application, but such tags are useful only in Facebook and are not really a part of the image file. The Facebook tags are more like folder names, as they are not really anchored to the underlying data.

When we annotate text files for use in linguistic analysis, semantics, business intelligence, enterprise search and other applications, we apply the concept of annotaions to specific ranges of text within a document. We do this for several possible reasons: to speed retrieval, for example, or to aid retrieval by providing a set of guide words drawn from the text, or to add guide words that the text does not already contain. Or, annotations can connect specific text in a document to broader concepts and background knowledge stored in a knowledge base, ontology or other external resource. The tree of connections that derive from annotations can be visualized like this:



We might also use annotations to help classify documents, or compare them with each other. Comparison could be used to aid in machine translation or to develop a common vocabulary. Or, we might be trying to extract only the bits of information we need to avoid having to read the entire document, or find examples in a document that are different from things we already know about but might be related, such as potentially new popular terms for things. Sometimes, we want to correct the original document without actually changing it. In short, there are many useful things we can do with annotations.

There are two basic processes for adding such data: automatic and manual. Automatic annotation is less precise but can operate over many more documents than humans can reasonably address. Manual is more precise (to a point), but very labor-intensive and is often use to train a machine to perform automatic annotation.

A good annotation-based system uses both—so-called "semi-automatic" annotation. In semi-automatic annotation, manual steps can come in several parts of the overall process. An initial manual step might identify a basic set of data or terms. These would be used to create a list of words that a computer could find in many documents (thousands or millions). Then a manual step would refine what was found by the computer and the results would be fed back into the automatic process to make it more precise. Later, a manual process might extend the automatic step for very specific uses, or to create new data as needs change.

For our purposes, we mark spans of text in some way using GATE Teamware, an online annotation management tool. We will associate specific spans of text with specific labels, often to be used later for some sort of processing. In most cases, manual annotation will follow strict guidelines describing what spans to annotate and how to label them. The following guidelines describe the general processes for annotation depending on the project type. The general guidelines will almost always be accompanied by precise guidelines specific to a project, including:

- What text to annotate,
- What labels to apply in what circumstances, and
- How to deal with special cases

Types of Annotation Projects

There are many types of annotation projects, some of which do not fit neatly into categories. Below are four examples that describe the most common types of projects. The general guidelines cover all of these projects, but the specific guideline processes may be quite different.

Gold Standard

The goal of the Gold Standard is to create a top threshold for measuring computer performance through manual annotations applied to the same text by more than one person. This is often used at the outset of a project to see how often two or more

annotators agree on a particular annotation or type of annotation. When the independent annotators reach some level of agreement, maybe 85% of the time, that becomes the target for a computer-generated annotation. In other contexts, we might created a very precise manual annotation of 100% accuracy in order to train an automatic process (or judge the effectiveness of an automatic process).

Quality Assurance

A quality assurance project often looks for errors in previously-processed documents. The errors may either be corrected, or simply collected and analyzed, depending on the need. A quality assurance process may also create its own gold standard on a subset of all documents—a mini corpus—to target specific problems. QA processes often seek 100% accuracy, unlike other processes, but on small, manageable sets of documents. Thus, the guidelines for annotators normally will be more strict and the process somewhat different.

Processing

A processing project may seek to add annotations that would be used in a processing step and then could be removed. For example, a processing project might highlight errors in a document and provide corrective labels which could be used to make automatic corrections in the original and then stripped from the text itself. Or, a processing annotation could be used in automatic routing of a document in an information management system. There are many types of processing projects. Typically, the labels will be quite specific with no possibility for variation.

Social

Social annotation (or "social tagging") is becoming increasingly popular both on the Web and within organizations. Often social tagging is an attempt to arrive at a new nomenclature or to accommodate bottom-up terminology (AKA "folksonomy") with a more formal system of names and labels. Social annotation might uncover new popular trends, or help with navigation, or generate civil discussion of various topics. In general, there is no attempt to constrain the annotations or how they are applied, but to harvest the various inputs to see if/when patterns emerge.

The Annotation Process

It would be possible to read a document from start to end, marking all annotations in order, as they are found. This has not, however, been found to give the most accurate results. In order for annotations to be consistent, a more methodical approach has to be applied, according to which **all** annotators should perform these actions in this particular order:

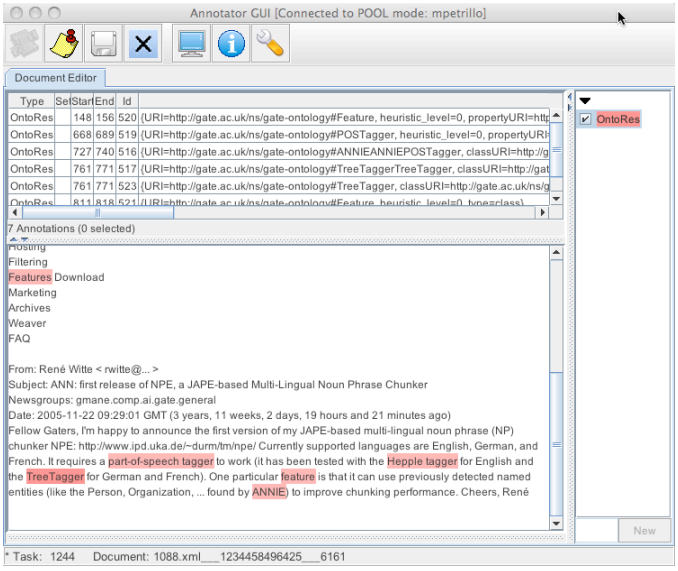
1. **Read the whole document.** Read the document through in its entirety, marking no annotations, to get an understanding.

2. **Mark the entities.** Read the document a second time, adding annotations for the mentions (including pronouns) of these basic entities, (in parallel if it is easier).

3. **Look again.** Review your work to be certain that you have not missed anything and especially that the annotation types and features are correct. Certain entities may suggest that other also exist, so also look for these if appropriate.

4. **Record any additional information.** As you are annotating the document, record any comments that you feel are important. You may have to do this in some text file/wiki, or perhaps in the annotation tool itself.

   For example, record:
   - Whether an annotation decision was hard to make.
   - Any questions or uncertainty you may have about the annotations and guidelines.
   - Anything unclear or ambiguous in the guideline.
   - Things that you consider important, which were not covered by the annotations allowed.
   - Annotation tool bugs and/or issues.



This is the annotation interface in GATE Teamware. It is the primary tool we use for manual annotation. Refer to the GATE Teamware Users Guide for instructions.

General Principles

The following are helpful principles to keep in mind.

- Annotation is not about trying to attach a label to every word.
- Annotation is about finding those things that are listed in the guidelines.
- Annotators must not collaborate when marking up texts, *unless explicitly requested to do so*.
- Every set of annotations should be the work of a single annotator only.

- There will be a designated phase of the annotation process for the discussion and resolution of differences between the work of multiple annotators.
- We are interested in consistent annotation based on a single, written set of instructions.
- Justify every annotation against the guidelines.
- Annotations should not overlap each other, or be contained within each other, unless the specific guidelines allow this.
- All annotations will be of a single document in isolation, and should not consider other documents.

Glossary of Terms

Annotation Meta-data added to a specific span of text.

Linguistic Annotation – Linguistic annotation goes beyond simple tagging to take into account the linguistic properties of a word or phrase, such as its part of speech, whether it is plural or singular, or whether it refers to another word in the surrounding text ("anaphora"). This information can be an important parameter when developing annotation guidelines. For example, whether a word is a noun or verb can affect how it is labeled.

Semantic Annotation – Semantic annotation connects a word or span of text to a semantic database or ontology where additional informationis stored. Semantic annotation transforms the target text into an "entity," which is a specific data element in a universe of data elements. A semantic database codifies additional information about such entities, including relationships between entities, into a format that can be processed by computers. Semantic Annotation also provides an anchor point in a text document for examples (or "instances") of such entities. Semantically annotated documents, therefore, can be connected to a wealth of searchable information useful in many information managment contexts. In Semantic Annotation, the traditional annotation type and features may be replaced by an address in the form of URI (Universal Resource Indicator).

Corpus/Corpora -- a corpus is a set of documents. *Corpora* is the plural. In Teamware, corpora are the heart of projects. Each project works on a specific corpus. Two projects cannot work on the same corpus simultaneously.

Entity – something that has a distinct, separate existence independent of the text. It can be either implicit (not mentioned, but its existence may be inferred from the text), or explicit (mentioned directly in the text). See also, *Semantic Annotation*.

Mention – A textual realization of an entity. In other words, a mention is how the entity is explicitly presented within the text. Implicit entities have no mentions.

Label – the word or phrase the describes an annotation. The label ties the annotation (and its mention in the text to an entity) A label has two parts:

Type – the generic word that describes an entity

Feature – more specific sub categories of the type

For example an annotation labeled "address" (the type) can be either "street address", "city" or "country" (the features).

Features may also be additional information, such as the mention's part of speech.

Set In GATE Teamware, we collect annotations into sets for further processing. Sets are normally associated with a particular annotator. This helps us compare annotations created be different annotators on the same document.

Inter-Annotator Agreement (IAA) This is a formal process for determining how well two annotators agree on a given annotation in the same document. GATE Teamware includes tools to automatically calculate an IAA value for a particular document. Other GATE tools can calculate IAA across an entire corpus.

Annotation DIFF An automatic process supported by GATE Teamware and other GATE tools to show how and where annotations differ within and between documents. It is similar to IAA but does not calculate a percentage of agreement.

Curate/Curator In text analysis, to curate is to collect the best examples and/or eliminate errors. In most cases, this process is indistinguishable from "editing." However, for some projects we may wish to keep several annotations, especially in cases where there is no way to determine "right" or "wrong." The curator often has specific knowledge of the contents of the document and can make an expert choice between two or more choices, or note whether one annotation is more correct than another.

Domain – the domain is the subject matter of the document in question. The concept of domain helps us identify context for the content. For example, if the document is in the "domain" of computer science, the word "root" is more likely to be relevant to a file system than a tree. Semantic annotation takes these relationships into account when applying annotations. We often use the terms "domain knowledge" or "domain expert" to describe annotators or curators who really "understand" the content of a given document.