Annotation

Eduard Hovy
Information Sciences Institute
University of Southern California
email: hovy@isi.edu

1. Introduction

As researchers seek to apply their machine learning algorithms to new problems, corpus annotation is increasingly gaining importance in the NLP community. But since the community currently has no general paradigm, no textbook that covers all the issues (though Wilcock's book published in Dec 2009 covers some basic ones very well), and no accepted standards, setting up and performing small-, medium-, and large-scale annotation projects remains something of an art.

To attend, no special expertise in computation or linguistics is required.

2. Content Overview

This tutorial is intended to provide the attendee with an in-depth look at the procedures, issues, and problems in corpus annotation, and highlights the pitfalls that the annotation manager should avoid. The tutorial first discusses why annotation is becoming increasingly relevant for NLP and how it fits into the generic NLP methodology of train-evaluate-apply. It then reviews currently available resources, services, and frameworks that support someone wishing to start an annotation project easily. This includes the QDAP annotation center, Amazon's Mechanical Turk, annotation facilities in GATE, and other resources such as UIMA. It then discusses the seven major open issues at the heart of annotation for which there are as yet no standard and fully satisfactory answers or methods. Each issue is described in detail and current practice is shown. The seven issues are: 1. How does one decide what specific phenomena to annotate? How does one adequately capture the theory behind the phenomenon/a and express it in simple annotation instructions? 2. How does one obtain a balanced corpus to annotate, and when is a corpus balanced (and representative)? 3. When hiring annotators, what characteristics are important? How does one ensure that they are adequately (but not over- or under-) trained? 4. How does one establish a simple, fast, and trustworthy annotation procedure? How and when does one apply measures to ensure that the procedure remains on track? How and where can active learning help? 5. What interface(s) are best for each type of problem, and what should one know to avoid? How can one ensure that the interfaces do not influence the annotation results? 6. How does one evaluate the results? What are the appropriate agreement measures? At which cutoff points should one redesign or re-do the annotations? 7. How should one formulate and store the results? When, and to whom, should one release the corpus? How should one report the annotation effort and results for best impact?

The notes include several pages of references and suggested readings.

3. Tutorial Overview

1. Toward a Science of Annotation
   a. What is Annotation, and Why do We Need It?
2. Setting up an Annotation Project
   a. The Basic Steps
   b. Useful Resources and Services
3. Examples of Annotation Projects
4. The Seven Questions of Annotation
   a. Instantiating the Theory
   b. Selecting the Corpus
   c. Designing the Annotation Interface
   d. Selecting and Training Annotators
   e. Specifying the Annotation Procedure
   f. Evaluation and Validation
   g. Distribution and Maintenance
5. Closing: The Future of Annotation in NLP

Tutorial Abstracts of ACL 2010, page 4,
Uppsala, Sweden, 11 July 2010.