TUTORIAL

CORPUS ANNOTATION: FRAMEWORK AND HANDS-ON EXERCISE

LREC 2008
Marrakesh, Morocco

Eduard Hovy
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
USA
hovy@isi.edu
http://www.isi.edu/~hovy

Julia Lavid
Departamento de Filologia Inglesa
Universidad Complutense de Madrid
28040 Madrid
Spain
lavid@filol.ucm.es
http://www.ucm.es/info/atg/webpages/lavid/julia-webpage.html

Corpus annotation has gained increasing interest in the NLP community over the past 5 years; there are now annotation projects in several major research centers. Nonetheless, setting up and performing a large-scale annotation project remain somewhat of an art. While certain aspects of it are becoming more systematized, the community has no general paradigm, no textbook, and no generally accepted standards a present. This presents something of a problem to people wanting to create their own training material for their projects. The increasing number of talks, workshops, and papers on and around the topic of annotation testify to the growing importance of this theme.

This tutorial provides the participant with an in-depth look at the basic issues and problems in corpus annotation, as well as hands-on experience through two annotation exercises.

In order to apply automated language processing technology to assist humans with analysis and other text-oriented tasks such as retrieval, summarization, question answering, and translation, the technology has to be ‘trained’ to the particulars of the domain and the analysis task(s). The training procedure involves preparing a selection of the representative texts to create what is called the *training suite*. Typically, domain experts view the texts with suitable interfaces and in various ways and formats enter information they find useful for their task(s), in a process called *coding* or *annotation*. Usually, annotation includes the steps of delimiting some fragment of text, selecting one or more interpretive labels to attach to that portion, and perhaps adding additional information. Once two or more annotators have performed coding on the same texts, and have achieved a high enough degree of agreement between them, the language processing technology can be trained on a portion of the training suite, and its performance measured on the remainder. If that is satisfactory, the technology can be applied to additional, unannotated, material of the same type, thereby assisting analysts in future tasks.

Annotation is not an exact science. To help ensure clean and trustable annotations suitable for machine learning, the language processing community is beginning to address a set of seven issues. Using examples from several of Dr. Hovy’s projects, this talk describes each issue, lists some relevant work for each, and points to what needs to be resolved. The seven issues are: 1. How does one obtain a balanced corpus to annotate, and when is a corpus balanced (and representative)? 2. How does one decide what specifically to annotate? How does one adequately capture the theory behind the phenomena and express it in simple annotation instructions? 3. When hiring annotators, what characteristics are important? How does one ensure that they are adequately (and not over- or under-) trained? 4. How does one establish a simple, fast, and trustworthy annotation procedure? What interfaces does one build? How does one ensure that the interfaces do not influence the annotation results? 5. How does one evaluate the results? What are the appropriate agreement measures? At which cutoff points should one redesign or re-do the annotations? 6. Hoe should one formulate and store the results? How does one ensure compatibility with other existing resources? How does one make results available for best impact? 7. How does one report the annotation effort and results? How does one actually publish papers on this work? What should the papers contain?

This tutorial will have two parts. In the first part, the above issues and seven questions will be described in detail, using as primary example the OntoNotes project, in which Dr. Hovy is a participant. This will prepare the tutorial participants for the practicalities of setting up and conducting an annotation project. In the second part, all participants will perform a real annotation exercise, focusing on discourse phenomena, in which Prof. Lavid is an expert. The exercise will be followed by a discussion of the problems encountered and by tallying the annotations results, in order to compute agreement scores and analyze their correspondence to the problems encountered during defining and performing the annotation exercise.

Outline

1. Toward Annotation Science
   a. What is Annotation, and Why do We Need It?
   b. An Example of Annotation: OntoNotes
   c. The Seven Open Questions of Annotation
2. Break
3. Hands-on Annotation Exercise
   a. Task Definition
   b. Exercise 1
   c. Evaluation and discussion
   d. Exercise 2
   e. Evaluation and discussion
4. Closing discussion