Quality Control of Corpus Annotation through Reliability Measures

Ron Artstein
Department of Computer Science
University of Essex, Wivenhoe Park
Colchester CO4 3SQ
United Kingdom
artstein@essex.ac.uk

Abstract

The need for quality control of corpus annotation should be obvious: research based on annotated corpora can only be as good as the annotations themselves. In recent years, corpus annotation has expanded from marking basic morphological and syntactic structure to many new kinds of linguistic phenomena. Each new annotation scheme and every individual set of annotation guidelines need to be checked for quality, because quality inferences do not carry over from one scheme to another. A standard way of assessing the quality of an annotation scheme and guidelines is to compare annotations of the same text by two or more independent annotators. While researchers are generally aware of this technique, it seems that the inner workings of the statistics involved and how to interpret them are understood by few. Mechanical application of agreement coefficients found in software packages can lead to serious errors, and it is therefore crucial for people who do research on annotation to become intimately familiar with these statistics.

This tutorial is a thorough introduction to the statistics used for measuring agreement between corpus annotators, and hence for inferring the reliability of the annotation. The tutorial will focus on the mathematics of the various agreement measures, and consequently on the implicit assumptions they make about annotators and annotation errors. A major part of the tutorial will be devoted to agreement coefficients of the kappa family, which are the most commonly used reliability measures in computational linguistics, but the tutorial will also discuss alternative measures such as latent class analysis. The tutorial will not assume advanced mathematical knowledge beyond basic probability theory, and will thus be accessible to most researchers in computational linguistics.

Tutorial outline

1. Motivation
   - Reliability as an indicator of annotation quality in the absence of a test for correctness
   - Agreement between coders as a measure of reliability
2. How to measure agreement
   - Correction for chance agreement (Scott's Pi)
   - Individual coder bias (Cohen's Kappa)
   - Multiple coders (Fleiss)
   - Weighted coefficients (Cohen's weighted Kappa, Krippendorff's Alpha)
3. How to interpret agreement coefficients
   - Using agreement coefficients to infer the proportion of difficult items (Aickin)
   - Inferring error rates on different classes of items (latent class models)
4. Using agreement measures
   - Identifying strong and weak parts of the annotation
   - Adapting the coefficients for specific uses