The Annotation Conundrum

Mark Liberman
University of Pennsylvania
myl@cis.upenn.edu

Abstract

Without lengthy, iterative refinement of guidelines, and equally lengthy and iterative training of annotators, the level of inter-subjective agreement on simple tasks of phonetic, phonological, syntactic, semantic, and pragmatic annotation is shockingly low.

This is a significant practical problem in speech and language technology, but it poses questions of interest to psychologists, philosophers of language, and theoretical linguists as well.

1 Introduction

Biologists believe that they know what genes, organisms, chemical compounds, and diseases are. Linguists believe that they know what nouns, verbs, and clauses are. Ordinary literate speakers of English believe that they know what people, places, and organizations are. And all of them believe that they can recognize and understand instances of these categories in coherent text.

When two biologists, two linguists, or two English speakers discuss such texts, it seems plausible that they have understood such instances in the same way. Nevertheless, if they are asked to highlight these instances, the level of inter-subjective agreement will be shockingly low.

Similarly depressing results are obtained in tasks such as phonetic or surface-phonemic transcription, co-reference annotation, identification of animacy, etc. Things are usually not much better if we compare annotations produced by the same individuals on different occasions.

A solution exists, in the practical sense of producing annotations with high inter-annotator agreement scores. The initially-divergent results of multiple annotations are discussed and adjudicated, and principles of interpretation are defined for future use. This process is repeated over and over again, typically for several months, until the desired level of agreement is obtained, or funding runs out.

At least for simple linguistic annotation tasks, this process, reminiscent of the development of common law, generally converges (though the residual level of disagreement may be depressingly high, especially when multiple judgments must be cascaded). The resulting annotation manuals may be hundreds of pages long, even for fairly limited tasks; and new annotators face weeks or months of training to become competent in learning to apply them.

There are several obvious ideas about why this might be true, but most of these ideas seem to be false. It will be argued that part of the answer lies in understanding that most linguistic annotation tasks are not really classification problems, but rather translation problems. We don't normally assume that there is only one correct translation of a Chinese sentence into English; nor do we try to make this true by constructing elaborate translation guidelines to cover every relevant contingency, though in principle we could.

Implications in engineering and science will be discussed.

Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics, page 2,
Athens, Greece, 30 March, 2009.