Automatic Time Expression Labeling for English and Chinese Text

Kadri Hacioglu, Ying Chen, Benjamin Douglas

Center for Spoken Language Research
University of Colorado at Boulder
Boulder, Colorado 80309

Abstract. In this paper, we describe systems for automatic labeling of time expressions occurring in English and Chinese text as specified in the ACE Temporal Expression Recognition and Normalization (TERN) task. We cast the chunking of text into time expressions as a tagging problem using a bracketed representation at token level, which takes into account embedded constructs. We adopted a left-to-right, token-by-token, discriminative, deterministic classification scheme to determine the tags for each token. A number of features are created from a predefined context centered at each token and augmented with decisions from a rule-based time expression tagger and/or a statistical time expression tagger trained on different type of text data, assuming they provide complementary information. We trained one-versus-all multi-class classifiers using support vector machines. We participated in the TERN 2004 recognition task and achieved competitive results.

1 Introduction

Extraction of temporal expressions from an input text is considered a very important step in several natural language processing tasks; namely, information extraction, question answering (QA), summarization etc. (Mani 2004). For example, in the summarization task, temporal expressions can be used to establish a time line for all events mentioned in multiple documents for a coherent summarization. Recently, there has been growing interest in addressing temporal questions in QA systems (Schilder and Habel 2003; Saquete et. al. 2004). In those systems, a highly accurate temporal expression recognizer or tagger (statistical or rule-based) is required for effective treatment of temporal questions yielding high-quality end-to-end system performance.

An official evaluation, sponsored by the DARPA automatic content extraction (ACE) program, has been organized by MITRE and NIST for time expression recognition and normalization (TERN) in 2004. The TERN task requires the recognition of a broad range of temporal expressions in the text and normalization of those expressions according to (Ferro et. al. 2004). The annotation is intended to mark information in the source text that mentions *when* something happened, *how long* something lasted, or *how often* something occurs. Temporal expressions in text vary from explicit references, e.g. *June 1, 1995*, to implicit