# Incremental Information Extraction Using Tree-based Context Representations

Christian Siefkes

Berlin-Brandenburg Graduate School in Distributed Information Systems*
Database and Information Systems Group, Freie Universität Berlin
Takustr. 9, 14195 Berlin, Germany
christian@siefkes.net

**Abstract.** The purpose of *information extraction* (IE) is to find desired pieces of information in natural language texts and store them in a form that is suitable for automatic processing. Providing annotated training data to adapt a trainable IE system to a new domain requires a considerable amount of work. To address this, we explore *incremental learning*. Here training documents are annotated sequentially by a user and immediately incorporated into the extraction model. Thus the system can support the user by proposing extractions based on the current extraction model, reducing the workload of the user over time.
We introduce an approach to modeling IE as a token classification task that allows incremental training. To provide sufficient information to the token classifiers, we use rich, tree-based context representations of each token as feature vectors. These representations make use of the heuristically deduced document structure in addition to linguistic and semantic information. We consider the resulting feature vectors as ordered and combine proximate features into more expressive joint features, called "Orthogonal Sparse Bigrams" (OSB). Our results indicate that this setup makes it possible to employ IE in an incremental fashion without a serious performance penalty.

## 1 Introduction

The purpose of *information extraction* (IE) is to find desired pieces of information in natural language texts and store them in a form that is suitable for automatic querying and processing. IE requires a predefined output representation (*target structure*) and only searches for facts that fit this representation. Simple target structures define just a number of *slots*. Each slot is filled with a string extracted from a text, e.g. a name or a date (*slot filler*).

To adapt an IE system to a new domain, it is necessary to either manually rewrite the rules used in the system (in case of *static* rule-based systems) or to provide annotated training data (in case of *trainable* systems). Manual rewriting of rules is a time-consuming and intricate task that must be done by experts

* This research is supported by the German Research Society (DFG grant no. GRK 316).