

# Advanced Machine Learning for the Construction and Adaptation of Information Extraction Tools

Applied Software Engineering Research Group



Computational Sciences & Engineering Division

## Problem Statement:

- At the foundation of many applications that perform analysis over sets of natural language texts lies the task of extracting information into a structured form. Despite some demonstrable successes, Information Extraction (IE) suffers from a major flaw in most real applications. The extraction task for which a tool was built is rarely identical to the task on which it is deployed, and shifting IE tools to new textual domains (e.g. from newswire to emails) results in significant performance drops, even for simple types of extraction and even for slight shifts in domain. The errors propagate through multiple subtasks resulting in even more significant performance reductions for more complex tasks. Modifying extraction systems to work on new domains or new tasks has traditionally been a tedious process and the cost was not always justifiable.

## Technical Approach:

- We utilize the latest machine learning (ML) models and expand on the growing success of ML methods that allow effective use of unlabeled data and that apply information learned from related tasks in order to allow efficient and effective porting of modern extraction tools to new information domains and to the extraction of new information types. We even extend these techniques to the building of extraction tools in other languages.

## Benefit:

- The customer can obtain tools that target the extraction of exactly the information that they require as opposed to a related category of meaning and that perform well on the specific types of data they need to analyze. Since nearly every type of analysis can be affected by errors inherent in the application of tools to a new domain and by the inability to target the precise class of meaning they want, these capabilities can improve a large list of potential applications that rely on automatic extraction of information from natural language.

Point of Contact:

Chris Symons  
(865) 241-5952  
symonsct@ornl.gov

