

Coupled Semi-Supervised Learning for Information Extraction

Andrew Carlson
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
acarlson@cs.cmu.edu

Justin Betteridge
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
jbetter@cs.cmu.edu

Richard C. Wang
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
rcwang@cs.cmu.edu

Estevam R. Hruschka Jr.
Federal University of Sao
Carlos
Sao Carlos, SP - Brazil
estevam@dc.ufscar.br

Tom M. Mitchell
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
tom.mitchell@cs.cmu.edu

ABSTRACT

We consider the problem of semi-supervised learning to extract categories (e.g., academic fields, athletes) and relations (e.g., PlaysSport(athlete, sport)) from web pages, starting with a handful of labeled training examples of each category or relation, plus hundreds of millions of unlabeled web documents. Semi-supervised training using only a few labeled examples is typically unreliable because the learning task is underconstrained. This paper pursues the thesis that much greater accuracy can be achieved by further constraining the learning task, by coupling the semi-supervised training of many extractors for different categories and relations. We characterize several ways in which the training of category and relation extractors can be coupled, and present experimental results demonstrating significantly improved accuracy as a result.

Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning—*knowledge acquisition*; I.2.7 [Artificial Intelligence]: Natural Language Processing—*text analysis*

General Terms

Algorithms, Experimentation

Keywords

Semi-supervised learning, bootstrap learning, information extraction, web mining

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WSDM'10, February 4–6, 2010, New York City, New York, USA.
Copyright 2010 ACM 978-1-60558-889-6/10/02 ...\$10.00.

1. INTRODUCTION

Machine learning approaches have been shown to be very useful for information extraction from text, including approaches that learn to extract various categories of entities (e.g., Athlete, City) and relations (e.g., CompanyProducesProduct) from structured and unstructured text [3, 28]. However, supervised training of accurate entity and relation extractors is costly, requiring a substantial number of labeled training examples for each type of entity and relation to be extracted. Because of this, many researchers have explored *semi-supervised* learning methods that use only a small number of labeled examples of the predicate to be extracted, along with a large volume of unlabeled text [5, 19, 1]. While such semi-supervised learning methods are promising, they often exhibit unacceptable accuracy because the limited number of initial labeled examples is insufficient to reliably constrain the learning process.

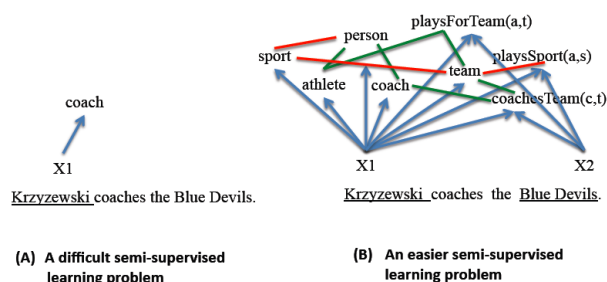


Figure 1: We show that significant improvements in accuracy result from coupling the training of information extractors for many interrelated categories and relations (B), compared with the simpler but much more difficult task of learning a single information extractor (A).

The thesis explored in this paper is that we can achieve much higher accuracy in semi-supervised learning by coupling the simultaneous training of *many* extractors, as suggested in Figure 1. The intuition here is that the underconstrained semi-supervised learning task can be made easier by adding new constraints that arise from coupling the training of many extractors.

