Available online at www.sciencedirect.com
SCIENCE DIRECT®
Artificial Intelligence 165 (2005) 91–134
Artificial Intelligence
www.elsevier.com/locate/artint

# Unsupervised named-entity extraction from the Web: An experimental study ☆

Oren Etzioni *, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, Alexander Yates

*Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350, USA*

Received 2 November 2004

Available online 14 April 2005

**Abstract**

The KNOWITALL system aims to automate the tedious process of extracting large collections of facts (e.g., names of scientists or politicians) from the Web in an unsupervised, domain-independent, and scalable manner. The paper presents an overview of KNOWITALL's novel architecture and design principles, emphasizing its distinctive ability to extract information without any hand-labeled training examples. In its first major run, KNOWITALL extracted over 50,000 class instances, but suggested a challenge: How can we improve KNOWITALL's recall and extraction rate without sacrificing precision?

This paper presents three distinct ways to address this challenge and evaluates their performance. *Pattern Learning* learns domain-specific extraction rules, which enable additional extractions. *Subclass Extraction* automatically identifies sub-classes in order to boost recall (e.g., "chemist" and "biologist" are identified as sub-classes of "scientist"). *List Extraction* locates lists of class instances, learns a "wrapper" for each list, and extracts elements of each list. Since each method bootstraps from KNOWITALL's domain-independent methods, the methods also obviate hand-labeled training examples. The paper reports on experiments, focused on building lists of named entities, that measure the relative efficacy of each method and demonstrate their synergy. In concert, our methods gave KNOWITALL a 4-fold to 8-fold increase in recall at precision of 0.90, and discovered over 10,000 cities missing from the Tipster Gazetteer.

☆ This is a substantially expanded version of our AAAI '04 paper.

* Corresponding author.
*E-mail address:* etzioni@cs.washington.edu (O. Etzioni).

0004-3702/$ – see front matter 
doi:10.1016/j.artint.2005.03.001

*Keywords:* Information Extraction; Pointwise mutual information; Unsupervised; Question answering

## 1. Introduction and motivation

Information Extraction is the task of automatically extracting knowledge from text. *Unsupervised* information extraction dispenses with hand-tagged training data. Because unsupervised extraction systems do not require human intervention, they can recursively discover new relations, attributes, and instances in a fully automated, scalable manner. This paper describes KNOWITALL, an unsupervised, domain-independent system that extracts information from the Web.

Collecting a large body of information by searching the Web can be a tedious, manual process. Consider, for example, compiling a comprehensive, international list of astronauts, politicians, or cities. Unless you find the "right" document or database, you are reduced to an error-prone, piecemeal search. One of KNOWITALL's goals is to address the problem of accumulating large collections of facts.

In our initial experiments with KNOWITALL, we have focused on a sub-problem of information extraction, building lists of named entities found on the Web, such as instances of the class `City` or the class `Film`. KNOWITALL is able to extract instances of relations, such as `capitalOf(City,Country)` or `starsIn(Actor,Film)`, but the focus of this paper is on extracting comprehensive lists of named entities.

KNOWITALL introduces a novel, generate-and-test architecture that extracts information in two stages. Inspired by Hearst [22], KNOWITALL utilizes a set of eight domain-independent extraction patterns to *generate* candidate facts.[1] For example, the generic pattern "NP1 such as NPList2" indicates that the head of each simple noun phrase (NP) in the list NPList2 is a member of the class named in NP1. By instantiating the pattern for the class `City`, KNOWITALL extracts three candidate cities from the sentence: "We provide tours to cities such as Paris, London, and Berlin".

Next, KNOWITALL automatically *tests* the plausibility of the candidate facts it extracts using *pointwise mutual information* (PMI) statistics computed by treating the Web as a massive corpus of text. Extending Turney's PMI-IR algorithm [43], KNOWITALL leverages existing Web search engines to compute these statistics efficiently.[2] Based on these PMI statistics, KNOWITALL associates a probability with every fact it extracts, enabling it to automatically manage the tradeoff between precision and recall. Since we cannot compute "true recall" on the Web, the paper uses the term "recall" to refer to the size of the set of facts extracted.

Etzioni [19] introduced the metaphor of an *Information Food Chain* where search engines are herbivores "grazing" on the Web and intelligent agents are *information carnivores*

[1] Hearst proposed a set of generic patterns that identify a hyponym relation between two noun phrases. Examples are the pattern "NP {,} such as NP" and the pattern "NP {,} and other NP".

[2] Turney measured the similarity of two term based on how often the terms appear in proximity to each other in Web search-engine indices.