Web-Based Open-Domain Information Extraction

Marius Paşca
Google Inc.
1600 Amphitheatre Parkway
Mountain View, California 94043
mars@google.com

ABSTRACT

This tutorial provides an overview of extraction methods developed in the area of Web-based open-domain information extraction, whose purpose is the acquisition of open-domain classes, instances and relations from Web text. The extraction methods operate over unstructured or semi-structured text. They take advantage of weak supervision provided in the form of seed examples or small amounts of annotated data, or draw upon knowledge already encoded within resources created strictly by experts or collaboratively by users. The tutorial teaches the audience about existing resources that include instances and relations; details of methods for extracting such data from structured and semi-structured text available on the Web; and strengths and limitations of resources extracted from text as part of recent literature, with applications in knowledge discovery and information retrieval.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation

Keywords

Information extraction, Web corpora, knowledge acquisition

1. BIOGRAPHY

Marius Paşca is a research scientist at Google. He graduated with a Ph.D. degree in Computer Science from Southern Methodist University, Dallas, Texas and an M.Sc. degree in Computer Science from Joseph Fourier University, Grenoble, France. He served on the program committees of ACL, IJCAI, WWW, SIGIR, HLT, EMNLP, NAACL and AAAI, including area co-chair positions at HTL-06, CIKM-08, EMNLP-09 and CIKM-11. Current research interests include factual information extraction from unstructured text and natural-language matching functions for information retrieval.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.