LD4IE - Linked Data for Information Extraction

Overview

The World Wide Web provides access to tens of billions of pages, mostly containing information that is largely unstructured and only intended for human readability. On the other hand, the LOD provide billions of pieces of information linked together and made available for automated processing. However, there is the lack of interconnection between the information in the Web pages and that in LOD. A number of initiatives, like RDFa (supported by W3C) or Microformats (used by schema.org and supported by major search engines) are trying to enable machines to make sense of the information contained in human readable pages by providing the ability to annotate webpage content with links into LOD.

This has created unprecedented opportunities for Web-scale Information Extraction. First, it is the first time in the history of IE that a very large-scale information resource (LOD) is available, covering a growing number of domains, to serve as learning materials. Linked Data offers a uniform approach to link resources uniquely identifiable by URIs. This creates a large knowledge base of entities and concepts, connected by semantic relations. Such resources can be valuable seed data for IE tasks. Furthermore, the annotated web pages can be considered as training data in the traditional machine learning paradigm. The use of an uncontrolled and constantly evolving, community provided set of independent Web resource for IE is totally untapped in the current state of the art. Second, IE methods can also enhance web page annotation, which creates further linked data.

However, powering Web-scale IE using LOD faces major challenges. The first requirement for Web-scale IE is discovering relevant learning materials on LOD. It is non-trivial due to the highly heterogeneous vocabularies used by data publishers. Users are often required to be familiar with the datasets, vocabularies, as well as query languages that data publishers use to expose their data. Unfortunately, considering the sheer size and the diversity of LOD, imposing such requirements on users is infeasible. On the other hand, it is known that the coverage of domains can be very imbalanced and for certain domains the data can be very sparse. For example, according to http://lod-cloud.net, the ”Media”, ”Life Science”, ”Geographic” and ”Publications” domains represent over 44% of the LOD. ”Cross-domain” datasets (e.g., encyclopedia-oriented) represent only 13%. However, the relation between the quantity of training data and learning accuracy follows a non-linear curve with diminishing returns. The majority of LOD are created automatically by converting legacy databases with limited or no human validation, thus errors are present. Similarly, community-provided resources and annotations can contain errors, imprecision, spam, or even deviations from standards. Also, very large and regular resources can be redundant, i.e. contain a large number of instances that contribute little to the learning task, causing model overfitting, while introducing considerable overhead.

Addressing these challenges requires multi-field collaborative research effort covering various topics such as modeling IE tasks with respect to LD; efficient, large scale, and robust learning algorithms able to scale and cope with noise; measures for assessing learning material quality, and methods for selecting and optimizing training seeds.

Organising Committee

Program Committee