Creating Large-scale Training and Test Corpora for Extracting Structured Data from the Web

Robert Meusel and Heiko Paulheim

University of Mannheim, Germany
Data and Web Science Group
{robert,heiko}@informatik.uni-mannheim.de

Abstract. For making the web of linked data grow, information extraction methods are a good alternative for manual dataset curation, since there is an abundance of semi-structured and unstructured information which can be harvested that way. At the same time, existing structured data sets can be used for training and evaluating such information extraction systems. In this paper, we introduce a method for creating training and test corpora from websites annotated with structured data. Using different classes in schema.org and websites annotated with Microdata, we show how training and test data can be curated at large scale and across various domains. Furthermore, we discuss how negative examples can be generated as well as open challenges and future directs for this kind of training data curation.

Keywords: Information Extraction, Linked Data, Benchmarking, Web Data Commons, Microdata, schema.org, Bootstrapping the Web of Data

1 Introduction

The web of linked data is constantly growing, from a small number of hand-curated datasets to around $1,000$ datasets [1, 11], many of which are created using heuristics and/or crowdsourcing. Since manual creation of datasets has its inherent scalability limitations, methods that automatically populate the web of linked data are a suitable means for its future growth.

Different methods for automatic population have been proposed. Open information extraction methods are *unconstrained* in the data they try to create, i.e., they do not use any predefined schema [3]. In contrast, *supervised* methods have been proposed that are trained using existing LOD datasets and applied to extract new facts, either by using the dataset as a training set for the extraction [2, 13], or by performing open information extraction first, and mapping the extracted facts to a given schema or ontology [4, 12] afterwards. In this paper, we discuss the creation of large-scale training and evaluation data sets for such supervised information extraction methods.

2 Dataset Creation

In the last years, more and more websites started making use of markup languages as Microdata, RDFa, or Microformats to annotate information on their pages. In 2014, over

17.3% of popular websites made use of at least one of those three markup languages, with schema.org and Microdata being among the most widely deployed standards [5]. Tools like *Any23*[1] are capable of extracting such annotated information from those web pages and returning them as RDF triples.

On of the largest, publicly available collections of such triples extracted from HTML pages is provided by the *Web Data Commons* project.[2] The triples were extracted by the project using Any23 and Web crawls curated by the *Common Crawl Foundation*,[3] which maintains one of the largest, publicly available Web crawl corpora. So far, the project offers four different datasets, gathered from crawls from 2010, 2012, 2013, and 2014, including all together over 50 billion triples. The latest dataset, including 20 billion triples, which were extracted from over half a billion HTML pages, contains large quantities of product, review, address, blog post, people, organization, event, and cooking recipe data [9]. The largest fraction of structured data, i.e., 58% of all triples and 54% of all entities, use the same schema, i.e., schema.org[4], and the HTML Microdata standard[5] for annotating data. At the same time, being promoted by major search engines, this format is the one whose deployment is growing the most rapidly [8–10].

Since both the original web page and the extracted RDF triples are publicly available, those pairs (i.e., a web page and the corresponding set of triples) can serve as training and test data for a supervised information extraction system.

As the ultimate goal of an information extraction system would be to extract such data from web pages *without* markup, the test set should consist of non-markup pages. However, for such pages, it would be very time-consuming to curate a reasonably sized gold standard. As an alternative, we use the original pages from the Common Crawl and remove the markup. This removal is done by erasing all Microdata attributes found in the HTML code.

In order to train and evaluate high precision extraction frameworks, negative examples are also useful, i.e., examples for pages that *do not contain* any information for a given class (e.g., person data). While this is hard to obtain negative examples without human inspection, we propose the use of an approximate approach here: given that a page is already annotated with Microdata and schema.org, we assume that the website creator has annotated *all* information which can be *potentially annotated* with the respective method. Thus, if a web page which contains Microdata does not contain annotations for a specific class, we assume that the page does not contain any information about instances of that class.

Figure 1 summarizes the creation of the data sets and the evaluation process.

3 Dataset Description

The datasets that we created for evaluation focus on five different classes in schema.org. The classes were chosen in a way such that (a) a good variety of domains is covered and

[1] `https://code.google.com/p/any23/`

[2] `http://webdatacommons.org/structureddata`

[3] `http://commoncrawl.org/`

[4] `http://schema.org/`

[5] `http://www.w3.org/TR/microdata/`

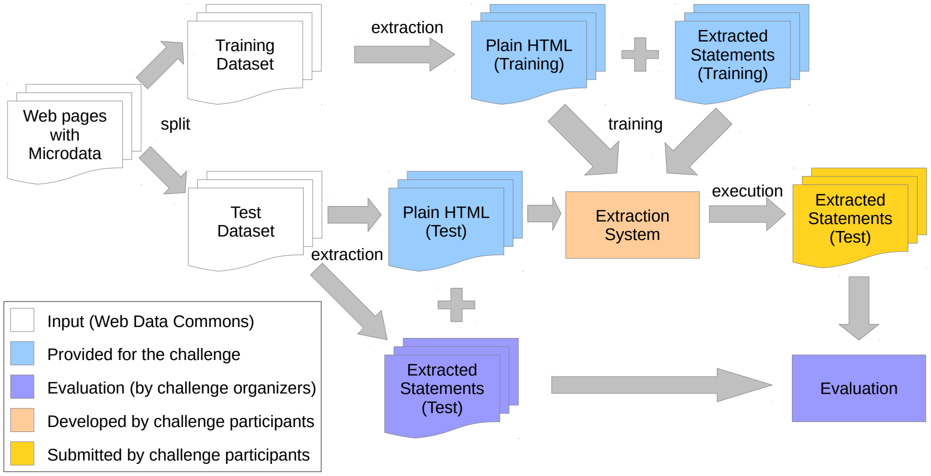
Web pages with Microdata
split
Training Dataset
extraction
Plain HTML (Training)
Extracted Statements (Training)
training
Test Dataset
extraction
Plain HTML (Test)
Extraction System
execution
Extracted Statements (Test)
Extracted Statements (Test)
Evaluation
Input (Web Data Commons)
Provided for the challenge
Evaluation (by challenge organizers)
Developed by challenge participants
Submitted by challenge participants

Fig. 1: Dataset creation and evaluation process

Table 1: Statistics about the training datasets provided

Class	Avg. instances per page	Avg. properties per page	# uniq. Hosts
MusicRecording	2.52	11.77	154
Person	1.56	7.71	2,970
Recipe	1.76	21.95	2,387
Restaurant	3.15	14.69	1,786
SportsEvent	4.00	14.28	135
Mixed	2.26	14.42	7,398

(b) the class is used by many different unique hosts. The latter is important, since for classes only deployed on a few different domains, which are potentially template-driven web sites, there is a danger of overfitting to those templates.

For each class, we provide a training dataset with minimal $7,000$ and maximal $20,000$ instances, and a test dataset with minimal $1,900$ and maximal $4,700$ instances.[6] Those can be used to set up systematic evaluations.[7]

In addition to the five class-specific datasets, we propose to evaluate approaches also on a mixed dataset, which contains instances from multiple classes. Table 1 shows some basic statistics about the datasets created.

4 Evaluation Metrics and Baselines

For evaluating information extraction systems that use the methodology described above in order to train models for information extraction, we propose to evaluate them using

[6] Note that for each page, there is exactly one root entity of the respective class, e.g., MusicRecording. The other entities are connected to the root entity, e.g., the artist and the record company of that recording.

[7] http://oak.dcs.shef.ac.uk/ld4ie2015/LD4IE2015/IE_challenge.html

Table 2: Minimal baseline for the datasets created. For the mixed class, one class was predicted at random.

Dataset	Recall	Precision	F-measure
MusicRecording	0.0690	1.0000	0.1291
Person	0.1105	1.0000	0.1990
Recipe	0.0430	1.0000	0.0825
Restaurant	0.0589	1.0000	0.1112
SportsEvent	0.0534	1.0000	0.1014
Mixed	0.0126	0.2071	0.0238

the originally extracted triples, using recall, precision, and F-measure as performance metrics. For obtaining stable results, the use of cross validation is advised.

The baseline for a class-specific extractor is creating a single blank node of the given schema.org class for each web page. This results in extractors of high precision (as the information is always correct) and low recall (since no further information is extracted). Such a system can be seen as a minimal baseline. Table 2 depicts the results of that baseline on the datasets discussed above.

For running challenges, such as the Linked Data for Information Extraction challenges [7], it is easily possible to create additional holdout sets, for which only the transformed web pages are given to the participants, while the corresponding original pages and triples are kept secret. This allows for participants to send in the triples they found and perform a comparison of different systems.

5 Conclusion

In this paper, we have shown that it is possible to create large-size training and evaluation data sets, which allows for benchmarking supervised information extraction systems. Using Microdata annotations with schema.org, we have discussed the creation of a corpus of training and test sets from various domains, ranging from recipes to sports events and music recordings. We have also discussed how to address the problem of generating negative examples.

While the corpus used in this paper focuses on schema.org and Microdata, similar datasets can be created when exploiting other markup languages, such as Microformats[8] [7] or RDFa[9]. Also, as existing crawl corpora might have limitations in terms of coverage, focused crawling for specific formats, vocabularies and classes can be applied to gather a sufficient data corpus for supervised learning as proposed in [6].

In the future, it will be interesting to see how existing information extraction systems perform given these datasets, as well as which new information extraction systems will be developed for bootstrapping the Web of Data.

[8] `http://microformats.org/`

[9] `http://www.w3.org/TR/xhtml-rdfa/`

References

1. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.
2. Fabio Ciravegna, Anna Lisa Gentile, and Ziqi Zhang. LODIE: linked open data for web-scale information extraction. In *Proceedings of the Workshop on Semantic Web and Information Extraction*, pages 11–22, 2012.
3. Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open information extraction: The second generation. In *IJCAI*, volume 11, pages 3–10, 2011.
4. Antonis Koukourikos, Vangelis Karkaletsis, and George A Vouros. Towards enriching linked open data via open information extraction. In *Workshop on Knowledge Discovery and Data Mining meets Linked Open Data (KnowLOD)*, pages 37–42, 2012.
5. Robert Meusel, Christian Bizer, and Heiko Paulheim. A web-scale study of the adoption and evolution of the schema. org vocabulary over time. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*, page 15. ACM, 2015.
6. Robert Meusel, Peter Mika, and Roi Blanco. Focused crawling for structured data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1039–1048. ACM, 2014.
7. Robert Meusel and Heiko Paulheim. Linked Data for Information Extraction Challenge 2014: Tasks and Results. In *Linked Data for Information Extraction*, 2014.
8. Robert Meusel and Heiko Paulheim. Heuristics for fixing common errors in deployed schema. org microdata. In *The Semantic Web. Latest Advances and New Domains*, pages 152–168. Springer, 2015.
9. Robert Meusel, Petar Petrovski, and Christian Bizer. The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In *13th Int. Semantic Web Conference (ISWC14)*, 2014.
10. Heiko Paulheim. What the adoption of schema. org tells about linked open data. In *2nd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '15)*, 2015.
11. Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the Linked Data Best Practices in Different Topical Domains. In *International Semantic Web Conference*, 2014.
12. Stephen Soderland, Brendan Roof, Bo Qin, Shi Xu, Oren Etzioni, et al. Adapting open information extraction to domain-specific relations. *AI Magazine*, 31(3):93–102, 2010.
13. Ziqi Zhang, Anna Lisa Gentile, and Isabelle Augenstein. Linked data as background knowledge for information extraction on the web. *SIGWEB Newsl.*, (Summer):5:1–5:9, July 2014.