Linked Data for Information Extraction Challenge 2014

Tasks and Results

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 to 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 Linked Data sets can be used for training and evaluating such information extraction systems. In this paper, we introduce the *Linked Data for Information Extraction Challenge 2014*. Using the example of person data in Microformats, we show how training and testing data can be curated at large scale. Furthermore, we discuss results achieved in the challenge, as well as open problems and future directions for the challenge.

Keywords: Information Extraction, Linked Data, Benchmarking, Web Data Commons, Microformats, 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, 7], many of them 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, 9], or by performing open information extraction first, and mapping the extracted facts to a given schema or ontology [4, 8]. In this paper, we discuss the creation of large-scale training and evaluation data sets for such supervised information extraction methods.

2 Task and Dataset

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 2013 over

13.8% of all websites made use of at least one of those three markup languages, where the most used markup format is Microformats hCard [5]. Tools like *Any23*[1] are capable of extracting such annotated information from those web pages and return them as RDF triples.

On of the largest, publicly available collections of such extracted triples 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 three different datasets, gathered from crawls from 2010, 2012 and 2013 including all together over 30 billion triples. The latest dataset, including 17 billion triples, which were extracted from over half a billion HTML pages, contain large quantities of product, review address, blog post, people, organization, event, and cooking recipe data [5].

Since both the original web page and the extracted RDF triples are publicly available, those pairs (a web page plus its corresponding triples) can serve as training data for a supervised information extraction system. In the 2014 edition of the challenge, we focus on one class of information only, i.e., Microformats data about persons using the *hCard* vocabulary.[4] For the challenge, we provide a training dataset both with and without markup, as well as a test set of web pages without the corresponding triples, which are kept as a non-public hold out set for evaluation. The training dataset consists of $9,877$ web pages and $373,501$ extracted triples, while the test dataset consists of $2,379$ web pages, with $85,248$ extracted triples (where the triples are not known to the challenge participants).[5]

Fig. 1 shows the distribution of predicates for both the training and the test set. It can be observed that the most frequent predicate is `rdf#type` (assigning the type `vcard#person`), followed by name attributes. There are no predicates which are exclusively contained in one of the two datasets.

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 substituting the Microformats classes with random strings, which are uniquely created for each web page. This is done to allow extraction systems to discover and exploit style information bound to those elements (e.g., person names displayed in bold).

In order to evaluate information extraction systems, participants were asked to send the extracted triples for the test set with the corresponding URL where the information was extracted from. We compared those to the original triples and computed recall, precision, and F-measure. Two triples from the URL are counted as identical if their subject, predicate, and object are all three identical URIs or literals, where blank nodes

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

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

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

[4] `http://microformats.org/wiki/hcard`

[5] The datasets, except for the triples of the test set, are available online at `http://data.dws.informatik.uni-mannheim.de/LD4IE/`

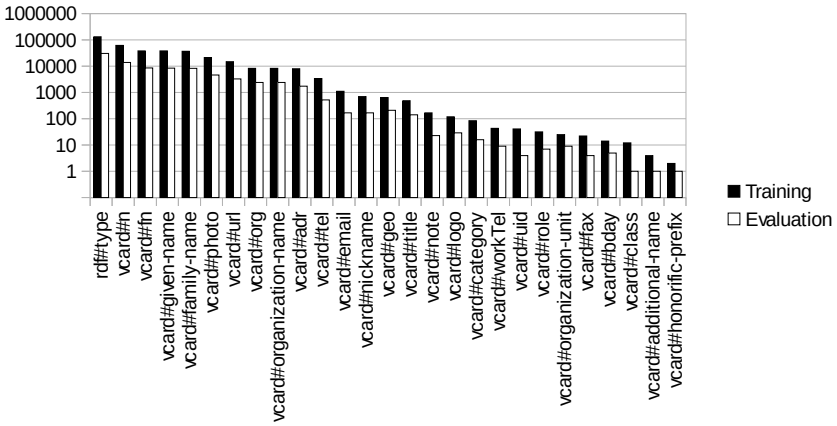

Fig. 1: Distribution of predicates in the training and test set

are always counted as identical.[6] Figure 2 summarizes the creation of the data sets and the evaluation process.

3 Results

One effect of the process for creating the 2014 datasets is that all web pages in the evaluation dataset can be expected to contain data about people, companies, or organizations. This allows for implementing a trivial baseline, i.e., creating a triple

```
_:1 rdf:type hcard:VCard .
```

for each web page.

We received one submission to the challenge, i.e., the μRaptor system [6]. In the following, we compare the baseline against that system, and provide some further insights in the results.

Table 1 shows the overall performance in terms of recall, precision, and F-measure. First of all, it is interesting that the baseline does not reach a precision of 1. This hints at pages for which the gold standard is not perfect: for example, names and other attributes of a person or organization are given, without explicitly stating the type `hcard#VCard`. In cases like these, a perfect information extraction system will not reach a precision of 1, based on the gold standard. One possible solution here is to use RDFS entailment[7] to materialize all axioms of both the gold standard and the solutions using RDFS inference on the vCard vocabulary.[8]

[6] The drawback of that convention is that for a web page containing n different `hcard#VCard` instances, a perfect solution correctly attributing all properties to n blank nodes cannot be distinguished from a solution attributing them all to one single blank node, although the latter is clearly inferior. Resolving that issue, however, requires graph matching with blank nodes, which is not a trivial problem, and such a solution might be prone to introducing other biases.

[7] `http://www.w3.org/TR/rdf11-mt/#rdfs-entailment`

[8] `http://www.w3.org/Submission/vcard-rdf/`

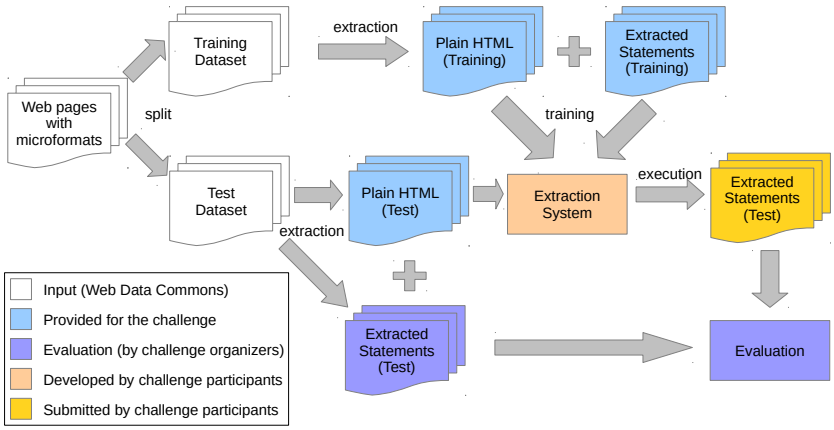

Fig. 2: Dataset creation and evaluation process

Table 1: Overall performance of the submitted system and the baseline.

Approach	# Triples	Recall	Precision	F-measure
μRaptor	61,909	0.665	0.916	0.771
Baseline	2,379	0.027	0.966	0.052

Another observation for the baseline is that, even for type statements, the recall of the baseline is low below 10%, since many pages contain data about more than one entity.

For the submitted solution μRaptor, we can observe a significant amount of information at a suitable precision. Fig. 3 depicts the performance of that system by predicate. A first observation is that more frequent predicates are more easily extracted (Pearson's correlation between frequency and F-measure is 0.687). While pictures are particularly well extracted (as they are rather easy to detect in HTML), organizations and addresses seem more difficult. The latter may be explained by our evaluation using string equivalence, which may not always be appropriate for complex organization names and addresses. On the other hand, it is interesting to see that the recall for telephone numbers is unusually low, compared to the other predicates.

4 Conclusion

This year, we initiated the first Linked Data for Information Extraction Challenge, showing that it is possible to create large-size training and evaluation data sets, which allows for benchmarking supervised information extraction systems. The task this year used hCard data, i.e., data about people, companies, and organizations.

The submitted results show that systems can be developed which, trained on a set of web pages, extract meaningful information. On the other hand, the results also show that

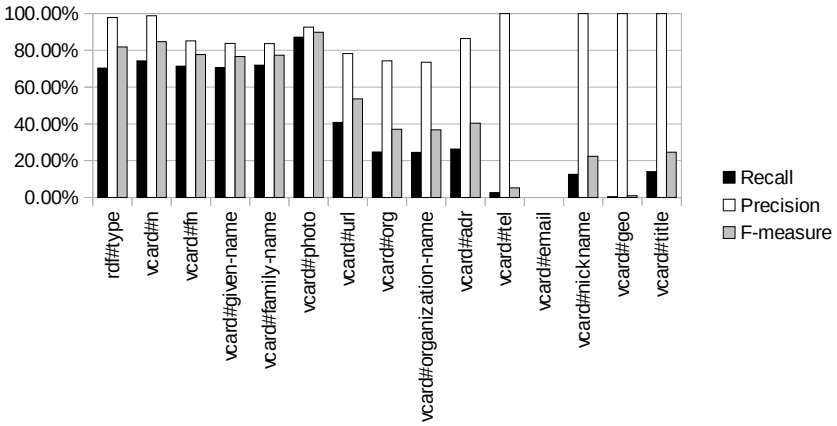

Fig. 3: Performance of μRaptor by Predicate

the proposed evaluation has some limitations by nature, compared by, e.g., a manually curated gold standard and supervised evaluation: not all the data in the gold standard may be complete and correct, and the strict comparison of extracted values may be too strict in some cases (e.g., when comparing addresses for string equality). Nevertheless, a manual creation and evaluation would not be scalable enough – which holds for the method proposed in this paper, which can be used to create training and test sets of nearly arbitrary sizes.

There are two main directions in which we want to extend this evaluation in the future. The first (and obvious) one is to include other classes as well, possibly also from different markup techniques (i.e., Microdata and RDFa).

The second direction is to make the task more realistic by mixing relevant and irrelevant pages. This year, the evaluation dataset was compiled from web pages that contain markup about persons. Thus, an extraction system could assume that some sort of person data could be found on that web page. In a more realistic setting, the extraction system would get a set of web pages which may or may not contain data of the desired type. Thus, future editions will foster two evaluation datasets per class: one with relevant pages (like this year), and one with a mix of relevant and irrelevant pages.

Since it is not reasonable to simply use a set of random, not marked-up web pages as irrelevant pages (as they may contain information of the desired type, but just no markup), one idea is to use marked-up web pages from which no data of the type at hand has been extracted. The rationale is that since the web page creator has used markup, it is likely that s/he would have also included markup for all other applicable types as well. Therefore, it is likely that the web page is irrelevant for the type at hand.

Another interesting question is how representative the data in the training sets is, which is relevant for the general applicability of systems trained based on that data. Since some wide-spread content management systems (CMS) create markup in web pages, it may be that the dataset shows a bias towards web pages created using those

CMS. In future editions of the challenge, we aim at a closer examination of such biases and blind spots.

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, Petar Petrovski, and Christian Bizer. The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In *13th Int. Semantic Web Conference (ISWC14)*, 2014.
6. Emir Muñoz, Luca Costabello, and Pierre-Yves Vandenbussche. μRaptor: A DOM-based system with appetite for hCard elements. In *2nd International Workshop on Linked Data for Information Extraction*, 2014.
7. Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the Linked Data Best Practices in Different Topical Domains. In *International Semantic Web Conference*, 2014.
8. 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.
9. Ziqi Zhang, Anna Lisa Gentile, and Isabelle Augenstein. "linked data as background knowledge for information extraction on the web" by ziqi zhang, anna lisa gentile and isabelle augenstein with martin vesely as coordinator. *SIGWEB Newsl.*, (Summer):5:1–5:9, July 2014.