Noname manuscript No.
(will be inserted by the editor)

Using the Semantic Web as a Source of Training Data

Christian Bizer · Anna Primpeli · Ralph Peeters

Received: April 12th, 2019 / Accepted: date

Abstract Deep neural networks are increasingly used for tasks such as entity resolution, sentiment analysis, and information extraction. As the methods are rather training data hungry, it is necessary to use large training sets in order to enable the methods to play their strengths.

Millions of websites have started to annotate structured data within HTML pages using the schema.org vocabulary. Popular types of entities that are annotated are products, reviews, events, people, hotels, and other local businesses [11]. These semantic annotations are used by all major search engines to display rich snippets in search results. This is also the main driver behind the wide-scale adoption of the annotation techniques.

This article explores the potential of using semantic annotations from large numbers of websites as training data for supervised entity resolution, sentiment analysis, and information extraction methods. After giving an overview of the types of structured data that are available on the Semantic Web, we focus on the task of product matching in e-commerce and explain how semantic annotations can be used to gather a large training dataset for product matching. The dataset consists of more than 20 million pairs of offers referring to the same products. The offers were extracted from 43 thousand e-shops, that provide schema.org annotations including some form of product identifiers, such as manufacturer part numbers (MPNs), global trade item numbers (GTINs), or stock keeping units (SKUs). The dataset, which we offer for public download, is orders of magnitude larger than the Walmart-Amazon [6], Amazon-Google [9], and Abt-Buy [9] datasets that are widely used to evaluate product matching methods. We verify the utility of the dataset as training data by using it to replicate the recent result of Mugdal et al. [14] stating that embeddings and RNNs outperform traditional symbolic matching methods on tasks involving less structured data. After the case study on product data matching, we focus on sentiment analysis and information extraction and discuss how semantic annotations from the Web can be used as training data within both tasks.

Keywords Entity resolution · Product matching · Sentiment analysis · Information extraction · Semantic Web · Schema.org annotations

Christian Bizer
University of Mannheim, Mannheim, Germany
Tel.: +49 621 181 2677
Fax: +49 621 181 2682
E-mail: chris@informatik.uni-mannheim.de

Anna Primpeli
University of Mannheim, Mannheim, Germany
Tel.: +49 621 181 2566
Fax: +49 621 181 2682
E-mail: anna@informatik.uni-mannheim.de

Ralph Peeters
University of Mannheim, Mannheim, Germany
Tel.: +49 621 181 2566
Fax: +49 621 181 2682
E-mail: rpeeters@mail.uni-mannheim.de

1 Semantic Annotations

Millions of websites have started to annotate data about products, people, organizations, places, local businesses, and events in their HTML pages using markup formats such as Microdata, JSON-LD, RDFa, and Microformats [11]. These annotations are used by all major search engines to display rich snippets in search results. The annotations are also one source of content of the knowledge graphs that are used by the search engines to rank search results and display knowledge cards


```
<div itemscope itemtype="http://schema.org/Product">
   <h1 itemprop="name">Signature Instant Dome 7 with integrated fly</h1>
   Item# <span itemprop="productID">2000015676</span>
   <img itemprop="image“ src=“../p_2000015676.jpg">
   <div itemprop="review" itemscope itemtype="http://schema.org/Review">
       <span itemprop="reviewRating" itemscope itemtype="http://schema.org/Rating">
       Rated <span itemprop="ratingValue"/>5 out of
       <span itemprop="bestRating"/>5</span> (a year ago)
       Review: <span itemprop="reviewBody">Great waterproof tent!</span>
   </div>
</div>
```

Fig. 1: Example of Microdata and schema.org annotations within an HTML page.

next to search results[1]. As Google, Bing, and Yandex recommend to use terms from the schema.org vocabulary,[2] this vocabulary is used by most websites. Figure 1 shows an example of how an offer and a review for a tent are annotated within an HTML page of an e-shop using the Microdata syntax[3] and the schema.org vocabulary. The left side of Figure 1 shows part of the HTML page as it is rendered by the browser. On the right, we see the corresponding source code. The *itemtype* attributes of the *DIV* and *SPAN* elements define the types of entities that are described, e.g. product and review. The itemprop attributes specify the properties that are used to describe the entites, e.g. *name*, *productID*, *image*, *ratingValue*, and *reviewBody*.

The Web Data Commons (WDC) project[4] monitors the adoption of schema.org annotations on the Web by analysing the Common Crawl[5], a series of public web corpora each containing several billion HTML pages [11]. The November 2018 version of the Common Crawl contains 2.5 billion pages originating from 32.8 million pay level domains (PLDs)[6]. Out of these PLDs, 9.6 million use semantic annotations (29.3%). Table 1 gives an overview of the most frequently offered types of data (schema.org classes). The table distinguishes between the two most widely used annotation syntaxes: The Microdata syntax for annotating data in the *BODY* of HTML pages and the JSON-LD syntax which is used to embed data into the *HEAD* section of HTML pages. As we see in the table, in sum around 850 thousand websites provide product data using the schema.org vocabulary. The product properties that are most widely used are *name*, *description*, *brand*, and *image*. Interestingly and very crucial for using semantic annotations from different websites to train matching methods, 30.5% of the websites annotate product identifiers, such as MPNs, GTINs, or SKUs, which allow offers for the same products to be clustered.

Table 1: Number of websites (PLDs) offering specific types of data.

schema.org class	#PLDs	
	Microdata	JSON-LD
WebPage	1,124,583	121,393
Product	812,205	40,169
Offer	676,899	57,756
Article	612,361	57,082
Organization	510,069	1,349,775
PostalAddress	502,615	178,500
ImageObject	360,875	111,946
BreadcrumbList	344,538	205,971
ListItem	338,845	209,207
Blog	337,843	12,174
BlogPosting	327,828	43,243
Person	324,349	335,784
LocalBusiness	294,390	249,017
AggregateRating	258,078	23,105
WebSite	158,054	3,519,466
Review	124,022	6,622
Place	92,127	66,396
Event	88,130	63,605
Brand	65,835	11,439

2 Cleansing Schema.org Product Data

Semantic annotations are placed in the templates, that are used to render HTML pages, by thousands of web masters. As these web masters have different levels of knowledge and different understandings of the schema.org vocabulary, schema.org terms are not used consistently and according to the specification on all sites. Thus, semantic annotations need to be cleaned before they can be used for training. In the following, we describe the pipeline of cleaning operations that we apply for creating our training dataset. We use the Web Data Com-

[1] https://developers.google.com/search/docs/guides/intro-structured-data
[2] https://schema.org/
[3] https://html.spec.whatwg.org/multipage/microdata.html
[4] http://www.webdatacommons.org/structureddata/
[5] http://commoncrawl.org/
[6] Examples of pay level domains are for instance amazon.de or ebay.co.uk

mons product corpus version November 2017[7] as starting point for the creation of the training set. The corpus contains 809 million *schema:Product* and *schema:Offer* entities originating from 581,482 websites. First, we select the subset of the offers that provide some kind of product identifier. Afterwards, we group the offers based on the identifiers into ID-clusters and cleanse abnormalities in the clustering. In the following, we provide details about both steps.

Selection of offers with identifiers. For the creation of the training corpus, we want to gather all products and offers which include a globally unique identifier and can this be clustered using this identifier. Examining the annotations, we notice that many websites annotate globally scoped identifiers, such as GTIN or MPN, using the vendor scoped term *sku* (stock keeping unit) or the generic terms *identifier* and *productID*. We thus consider all offers that contain any identifier-related term (e.g. *gtin8, gtin12, gtin13, gtin14, mpn, sku, identifier,* and *productID*) and try to filter out vendor-specific identifiers later in the cleansing process. Similar to the observations of [12] and [7], we notice that 6% of the websites annotating product offers have syntax errors in the URIs identifying schema.org terms or use deprecated or even undefined terms. As we do not want to miss these offers, we include all entities into our training set which have at least one property with an identity revealing suffix[8]. Using this selection strategy we find 116 million out of the 809 million offers (14%) in the Web Data Commons product corpus to contain some sort of identifier.

Detection and removal of listing pages and advertisements. We want to include the comprehensive description of a product from its detail page into the training set and not the summary of this information often found on listing pages and in advertisements on other detail pages. However, identifiers in listing items and advertisements are annotated as well which makes it necessary for us to detect those entities and remove them from the corpus. For the detection of listing pages and advertisements we use a heuristic which relies on the following features: amount of *schema.org/Offer* and *schema.org/Product* entities per web page, variation of the length of the product descriptions, number of identifier values, and semantic connection to parent entities using the terms *schema:relatedTo* and *schema:similarTo*. Our heuristic for identifying listing pages and advertisements achieves an F1 score of 94.8% on a manually annotated test set. This cleansing step removes 49% of the offer entities, leaving 58 million non-listing and non-advertisement offers in the training set.

Filtering by identifier length. In the next step, the identifier values are normalized by removing non-alphanumeric characters and common prefixes such as initial zero digits and identifier-related strings like *ean, mpn, sku,* and *isbn*. Considering the length of global identifiers such as GTIN or ISBN numbers in comparison to vendor-specific identifiers that are often relatively short, we filter out all offers having identifiers that are shorter than 8 characters. Additionally, offers whose id values completely consist of alphabetical characters are removed. Finally, we observe that a considerable number of websites use the same identifier value to annotate all their offers, likely due to an error in the script generating the pages. We detect these websites and remove their offers from the training set. After applying both filtering steps 26 million offer entities remain in the training set.

Cluster creation. We group the remaining 26 million offers into 18 million clusters using their ID values. It happens that single offers contain multiple alternative identifiers referring to the same product, e.g. GTIN8 and GTIN12, or GTIN12 and MPN. We use this information to merge clusters referring to the same product which results in a reduction of the number of clusters to 16 million. 13 million of these clusters contain only a single offer. We also notice that some websites include identifiers referring to product categories, such as UNSPSC numbers[9], in addition to identifiers referring to single products into the annotations. For detecting such cases, we examine the structure of the identifier co-occurrence graph within each cluster. We discover that vertices having a degree larger than 10 and a clustering coefficient of $C_i < 0.2$ tend to represent product categories rather than single products and we split the clusters accordingly. This leads to the creation of 199,139 additional clusters.

Offer categorization. The schema.org vocabulary contains terms for annotating the product category of an offer. However, less than 2% of the offer pages in the WDC 2017 corpus annotate category information. Different shops use different categorization schemata for presenting their products to the customers. We do not attempt to solve the resulting large-scale taxonomy integration problem, but re-classify the offers into 26 product categories that we selected from the upper parts of the Amazon product taxonomy. We use a pub-

[7] http://www.webdatacommons.org/structureddata/2017-12/stats/schema_org_subsets.html

[8] Regex applied to each predicate URI for capturing identity revealing properties:
$.*/(gtin8|gtin12|gtin13|gtin14|sku|mpn|identifier|productID)$

[9] http://www.unspsc.org/

licly available Amazon product and reviews dataset[10] and apply transfer learning [16] in order to assign product category labels to the clusters of our corpus. In cases for which the confidence of assigning a category label is low, we assign the label *other category*.

3 Profile of the WDC Training Dataset for Large-Scale Product Matching

Applying the cleansing procedure described above to the Web Data Commons product corpus (November 2017) results in a training data set consisting of 26 million offers originating from 79 thousand websites. The dataset has a compressed size of 6.4GB. We call the dataset *WDC Training Dataset for Large-Scale Product Matching* (WDC - LSPM). Using the identifiers, the offers are grouped into 16 million clusters referring to the same products. 13 million of these have a size of one, 1.9 have a size of two, and 1.1 million have a size larger than two. We also create an English-language subset which includes only offers from the top level domains *com*, *net*, *co.uk*, and *org*. The English language subset has a size of 3.9GB (compressed) and consists of 16 million offers which are grouped into 10 million clusters. Out of these clusters, 8.4 million have a size of one, 1 million have a size of two and 625.7 thousand have a size larger than two. Only considering clusters of English offers having a size larger than five and excluding clusters of sizes bigger than 80 offers which may introduce noise, 20.7 million positive training examples (pairs of matching product offers) and a maximum of 2.6 trillion negative training examples can be derived from the data set.

We extract all descriptive properties of offers that are annotated with schema.org terms. Table 2 shows the distribution of descriptive schema.org properties in the Full and English training sets as well as the distribution of identifier-related schema.org properties in both sets. We see that the density of the descriptive properties beside name and description is rather low ($< 50\%$). This is inline with earlier findings [11] that only a rather small subset of the schema.org vocabulary is actually widely used on the Web. We also see that over 75% of the identifiers that were used for clustering were annotated using the terms *sku* and *productID* which justifies our decision to not ignore these in theory vendor-specific properties but also consider their values in the cleansing process.

In addition to the annotated properties, we also extract product specifications in the form of key/value pairs from HTML tables that are included in the detail pages. For this we use the method described in the works of Petrovski et al. [15] and Qiu et al. [18]. The method detects specification tables for 24% of the offers contained in the Full set and 17% of the offers in the English set.

Table 3 shows the distribution of offers per product category as well as the clusters size distribution in the English training set. Considering clusters having a size larger than two, we can derive more than 1.2 million positive pairs for the clusters in the categories *Office Product* and *Clothing*, and more than 300 thousand pairs each for the categories *Shoes, Camera and Photo, Cell Phone and Accessories, Computers and Accessories,* and *Jewelry*. These amounts of pairs of offers referring to the same products are likely big enough for training even very data hungry entity matching methods and are even within each category alone much larger than the product matching datasets that have been available to the public so far (see Section 7).

4 Quality of the Clustering

In order to get an impression of the quality of the ID-clustering, we randomly sample 900 pairs of offers belonging to the same clusters and manually verify if the offers really refer to the same product by inspecting the *name* and *description* values of the offers. We discover that 93.4% of the pairs are correct, meaning that both offers refer to the same product. We find that 2.1% of the pairs in the sample (19 out of 900) are wrong due to web pages providing wrong identifier values. We consider this value to be low enough in order to use the identifiers for generating training pairs. We verify this assumption using the experiments described in Section 6. We further find that 1.0% of the pairs in the sample (9 out of 900) are wrong due to errors introduced by our transitive grouping strategy which combines two clusters if a single offer is found that is annotated with the identifier values of both clusters (e.g. a GTIN8 and a MPN number). In future work, we plan to investigate stricter merging criteria which might result in a better compromise between cluster size and amount of errors. For 3.4% of the sample (31 offer pairs), the authors of this article together were not able to decide whether the two offers refer to the same product as the names and descriptions were too short (e.g. just *"Samsung Galaxy"*) or too general (e.g. *"computer software"*). In 20 out of these 31 cases, name and description together contained less than four tokens. If desired, these pairs can be deleted from the training set using a length filter.

[10] http://jmcauley.ucsd.edu/data/amazon/

Table 2: Amount of offers in the training set and gold standard having specific properties.

Property	Offers Full set		Offers English Set		Gold Standard	
	#	%	#	%	#	%
s:name	25,281,317	95.37%	15,653,878	95.15%	2,095	99.61%
s:description	17,215,475	64.94%	11,352,319	69.00%	1,739	82.69%
s:brand	9,313,258	35.13%	5,645,282	34.31%	701	33.33%
s:image	5,785,250	21.82%	4,348,830	26.43%	321	15.26%
s:price	3,335,306	12.58%	1,977,269	12.01%	292	13.88%
s:priceCurrency	2,971,417	11.20%	1,873,611	11.38%	283	13.45%
s:availability	1,180,268	4.45%	716,066	4.35%	169	8.03%
s:manufacturer	2,024,537	7.63%	1,254,601	7.62%	285	13.55%
s:sku	11,475,859	43.29%	7,239,039	44.00%	671	31.90%
s:mpn	4,611,908	17.39%	3,167,895	19.25%	1,339	63.67%
s:productID	9,386,433	35.41%	6,351,147	38.60%	122	5.80%
s:gtin8	452,151	1.70%	167,411	1.01%	16	0.76%
s:gtin13	3,529,958	13.31%	1,449,759	8.81%	222	10.55%
s:gtin12	300,102	1.13%	261,431	1.58%	9	0.42%
s:gtin14	420.712	1.58%	71,428	0.43%	341	16.21%
s:identifier	179.225	0,67%	65,736	0.39%	5	0.23%

Table 3: Distribution of product categories in the English training set.

Product Category	%Offers	%ID-Clusters	#ID-Clusters of Size				
			2	[3-4]	[5-10]	[11-20]	[>20]
Office	12.02%	10.90%	99,237	40,314	16,920	5,953	3,043
Jewelry	7.90%	7.79%	94,389	33,156	13,329	3,352	2,037
Clothing	7.52%	6.80%	95,006	49,085	30,384	3,285	1,866
Automotive	5.40%	5.90%	54,320	16,650	8,139	2,865	2,140
Beauty	5.50%	5.78%	70,984	27,636	10,568	2,115	1,070
Cell Phones & Acc.	5.74%	4.76%	48,659	15,870	5,162	1,085	878
Home & Kitchen	7.23%	4.68%	45,922	24,429	7,414	1,247	538
Luggage & Travel	4.06%	4.48%	36,211	14,401	6,399	1,198	957
Tools	3.67%	4.35%	28,236	12,033	4,407	1,248	1,042
CDs & Vinyl	3.88%	4.19%	42,261	17,666	6,013	1,417	663
Shoes	3.93%	4.11%	37,647	16,603	7,590	1,335	721
Camera & Photo	3.23%	3.47%	39,924	14,583	5,408	1,423	935
Grocery	3.39%	3.26%	39,372	17,109	5,889	2,154	716
Computers & Acc.	4.13%	3.20%	48,125	11,614	5,411	2,308	2,862
Digital Music	2.58%	3.03%	23,678	7,954	3,046	640	535
Other Electronics	2.43%	2.83%	31,034	11,649	4,412	977	427
Books	2.19%	2.81%	22,350	9,889	2,946	330	183
Video Games	2.43%	2.62%	18,918	8,256	3,419	938	779
Garden	1.82%	2.43%	13,360	4,898	1,764	475	366
Musical Instruments	2.02%	2.31%	22,263	5,182	1,684	550	486
Pet Supplies	1.83%	2.15%	13,589	7,605	2,974	620	440
Baby	1.49%	1.71%	16,853	5,509	1,894	458	254
Toys	0.95%	1.19%	10,525	3,120	1,016	258	189
Sports	0.86%	0.75%	8,216	3,460	1,234	314	372
Movies & TV	0.62%	0.71%	6,588	2,030	681	195	157
Health	0.65%	0.70%	7,261	6,857	1,165	189	113
Other Category	2.54%	3.07%	37,292	12,964	4,088	633	657
TOTAL	100%	100%	1,012,220	400,522	163,356	37,562	24,345

5 WDC Gold Standard for Large-Scale Product Matching

Having some noise in the training set is acceptable, but should be avoided for the test set. We thus create a clean evaluation gold standard by manually verifying for 2,000 pairs of offers whether they refer to the same product or not. The 2,000 pairs of offers differ from the 900 pairs that we verified in order to assess the quality of the clustering. The level of difficulty of a matching task as well as the suitability of a matching method for the task both depend on the structuredness of the data to be matched. Thus, we select for the gold standard two product categories containing less structured offers (watches and sneaker shoes) as well as two categories containing more structured offers (computers & acces-

sories and camera & photo). First, we identify the clusters belonging to the selected product categories. We then sample positive pairs from within these clusters as well as textually similar negative pairs across clusters and manually check the correctness of the label. The resulting gold standard consists of 150 positive and 350 negative pairs for each category. The offers contained in the gold standard originate from the following numbers of clusters from each category: 338 for computers & accessories, 231 for camera & photo, 269 for watches and 186 for sneakers. The two right-most columns in Table 2 contain the density of the schema.org properties of the offers in the gold standard. The training set and the gold standard are provided for public download on the Web Data Commons website[11] which also provides additional statistics about both.

6 Entity Resolution Experiments

As a result of the success of embeddings and deep neural networks for tasks such as image recognition and natural language processing, the question whether these techniques also increase the performance of entity matching has recently moved into the research focus [4,19, 20,14]. Current results by Mudgal et al. [14] suggest that deep learning techniques perform comparable to traditional symbolic matching techniques on strongly structured data but outperform traditional techniques by a margin of 5% to 10% in F1 on less structured data such as product descriptions in e-commerce. The problem with these results is that they are not verifiable as they have been produced using training data *"from a major retailer"* [14] which is not available to the public.

This Section presents a set of matching experiments conducted using the English Training Set and the WDC gold standard. The experiments are intended on the one hand to verify the utility of our training set. On the other hand, we use our training set and gold standard to replicate the results of Mudgal et al. [14]. First, we perform an unsupervised bag-of-words (BOW) experiment using TF-IDF and cosine similarity. Afterwards, we train various supervised models such as logistic regression, naive Bayes, LinearSVC, decision trees, and random forests using (i) binary word co-occurrence vectors and (ii) string similarity scores, automatically generated by the Magellan framework [8], as features. As neural network based matchers, we combine all network types implemented in the *deepmatcher* framework (e.g. RNNs, Attention, and Hybrid, all with default parameters) with pre-trained and self-trained *fastText* embeddings.

We experiment with different subsets of the offer features *title*, *description*, *brand*, and specification table content. All identifier related properties (lower part of the Table 2) are removed from the offers. Due to resource limitations, we do not use the complete English training set for the supervised experiments but subsets of potentially interesting training examples (e.g. positive pairs from many different clusters and negative pairs from different clusters where both offers have a similar description). For the category computers, we use 20 thousand positive and 21 thousand negative training examples, for cameras 11 thousand positive and negative examples, for watches 6,289 positives and 9,161 negatives, and for sneakers 3,709 positives and 6,060 negatives.

The results of all experiments are summarized in Table 4. For each category, we report the best performing method/feature combination. As expected the supervised methods outperform the unsupervised BOW approach significantly. More interestingly, the deep learning approaches using pre-trained fastText embeddings are 8-10% better in F1 compared to the supervised methods using symbolic features. This confirms the result of Mudgal et al. that deep learning based matching methods excel on tasks involving less structured entity descriptions. More information about the exact configuration of all methods as well as the results of the not so good performing method/feature combinations are found on the project's web page.[12]

7 Comparison to Existing Entity Resolution Benchmark Datasets

Entity resolution is a long standing research area in which various benchmark datasets are used to compare matching methods. Table 5 gives an overview of entity resolution benchmark datasets along the dimensions: Public availability, number of sources from which the data originates, and number of positive pairs (e.g. records referring to the same real-world entity).

The two classic datasets in the area of product matching are *Abt-Buy* and *Amazon-Google* introduced by Köpcke, Thor, and Rahm [9]. Gokhale et al. introduce another public product dataset *Walmart-Amazon* [6]. In our previous work [17], we publish a gold standard for product data extraction and matching covering 32 different e-shops. Several datasets for evaluating duplicate detection methods are provided for public download by Naumann et al.[13]. The datasets describe movies,

[11] http://webdatacommons.org/largescaleproductcorpus/

[12] http://data.dws.informatik.uni-mannheim.de/largescaleproductcorpus/ExtResults.xlsx

[13] https://hpi.de/naumann/projects/repeatability/datasets.html

Table 4: Results of the product matching experiments.

Category	Classifier	Features	P	R	F1
Unsupervised Matching					
Computers	Cosine, TF-IDF, thr:0.25	title	0.50	0.89	0.64
Cameras	Cosine, TF-IDF, thr:0.3	title+desc	0.59	0.71	0.64
Watches	Cosine, TF-IDF, thr:0.35	title	0.48	0.91	0.63
Shoes	Cosine, TF-IDF, thr:0.4	title	0.57	0.80	0.66
Supervised Matching - Symbolic Feature Repr. - Value Co-Occurrence					
Computers	LinearSVM	title+desc+brand+spec	0.78	0.90	0.83
Cameras	LinearSVM	title+desc+brand	0.74	0.87	0.80
Watches	LinearSVM	title+desc+brand+spec	0.80	0.90	0.85
Shoes	LinearSVM	title+desc+brand	0.68	0.95	0.80
Computers	RandomForest	title+desc+brand+spec	0.76	0.88	0.81
Cameras	RandomForest	title+desc	0.80	0.83	0.82
Watches	RandomForest	title	0.77	0.87	0.83
Shoes	RandomForest	title+desc+brand+spec	0.70	0.88	0.78
Supervised Matching - Symbolic Feature Repr. - Magellan Feature Generation					
Computers	RandomForest	title+desc+brand+spec	0.65	0.85	0.74
Cameras	RandomForest	title+desc+brand+spec	0.61	0.83	0.70
Watches	RandomForest	title+desc+brand+spec	0.80	0.85	0.82
Shoes	RandomForest	title+desc+brand+spec	0.77	0.83	0.80
Supervised Matching - Pre-trained fastText Embeddings - DeepMatcher					
Computers	RNN	title+desc+brand+spec	0.89	0.95	**0.92**
Cameras	RNN	title+desc+brand+spec	0.90	0.95	**0.92**
Watches	RNN	title+desc+brand+spec	0.89	0.94	**0.91**
Shoes	RNN	title+desc+brand+spec	0.83	0.94	**0.88**

CDs, restaurants, scientific papers, and countries. Further benchmark datasets have been introduced for the Instance Matching Track of the Ontology Alignment Evaluation Initiative (OAEI)[14]. Daskalaki et al. give an overview of these datasets [3]. A large citation data set *Citeseer - DBLP* offering 550 thousand matches is provided in the Magellan Data Repository [10]. Finally, a large song data set containing 1.2 million matching pairs has been used to evaluate Falcon [2]. Mudgal et al. [14] use several large product datasets with up to 111 thousand positive pairs for evaluating their deep learning methods. Unfortunately, these datasets are not public.

The table shows that concerning the number of positive pairs our training datasets (WDC - LSPM and WDC - LSPM English) are four orders of magnitude larger than the other public evaluation datasets in the area of product matching. Compared to the Falcon-Songs data set, WDC - LSPM English is 17 times larger. Concerning the number of sources, WDC - LSPM English covers 43,293 sources while the existing datasets cover at most 32 sources. The other datasets do not explicitly distinguish between training and test set but leave the split to the user. We distinguish between training set and gold standard and give different quality guarantees for both.

[14] http://oaei.ontologymatching.org/

Table 5: Overview of entity resolution benchmark datasets.

Dataset	Publicly Available	# Data Sources	# Positive Pairs
Walmart-Amazon [6]	yes	2	1,154
Amazon-Google [9]	yes	2	1,300
Abt-Buy [9]	yes	2	1,097
DBLP-ACM [9]	yes	2	2,224
DBLP-Scholar [9]	yes	2	5,347
DM-Clothing [14]	no	1	105,608
DM-Electronics [14]	no	1	98,401
DM-Home [14]	no	1	111,714
DM-Tools [14]	no	1	96,836
DM-Company [14]	yes	?	28,200
OAEI - SYNTHETIC [1]	yes	1	1,800
OAEI - DOREMUS	yes	1	15
Citeceer - DBLP [10]	yes	2	558,787
Falcon - Songs [2]	yes	1	1,292,023
WDC - Product GS [17]	yes	32	1,500
WDC - LSPM	yes	79,126	40,582,671
WDC - LSPM English	yes	43,293	20,773,304

8 Using Semantic Annotations as Training Data for Other Tasks

The previous chapters have demonstrated the utility of semantic annotations for creating training data for product matching. Beside of product matching, semantic annotations can also be used to create large training sets for other tasks, such as information extraction or sentiment analysis. In this section we will discuss the potential of using semantic annotations within these two areas.

Information Extraction. Semantic annotations about types (e.g. product, event, hotel, local business, cooking recipe) and properties (e.g. name, address, opening hours, ingredient) together with structure of the HTML code around the annotations can be used to train information extraction methods to recognize the same type of information in web pages that do not contain such annotations. For instance, the annotation of the product price *69,99 Euro* within an HTML page provides the learning algorithm with an example of the structure and unit of measurement of price values as well as an example of the HTML structures that are used around price values on this page.

A successful example of an information extraction system that employs schema.org annotations as training data is the work of Foley et al. [5]. The purpose of their system is to discover data about local events, such as small venue concerts, theatre performances, garage sales, movie screenings, on web pages. To train their system they use event data from web pages which is annotated using the schema.org event properties *name*, *date*, *time*, and *location*. They evaluate their method on 700 million web pages from the ClueWeb12 corpus. Using 217,000 explicitly annotated events as supervision, they are able to double recall at a precision level of 85%. Unfortunately, they neither publish their code nor the event data set that they have extracted from the ClueWeb12 corpus.

A series of information extraction evaluation datasets that were built using schema.org annotations and which are public was compiled by Meusel and Paulheim for the information extraction challenge conducted at the Linked Data for Information Extraction (LD4IE) workshop 2014 and 2015. The dataset of the LD4IE Challenge 2014[15] consists of web pages containing Hcard[16] annotations describing contact information of persons and organizations. The goal of the challenge is to extract such contact information from pages without annotations. The dataset of the LD4IE Challenge 2015 [13][17] consists of HTML pages that contain schema.org annotations describing music recordings, persons, cooking recipes, restaurants, and sports events. This data set was extracted from the December 2014 version of the Common Crawl. Altogether, the pages originate from 7,300 different websites. The goal of the challenge is to extract such information from pages without annotations.

Sentiment Analysis. The goal of sentiment analysis is to determine the polarity of a given text towards an entity or different aspects characterizing the entity [21]. State of the art sentiment detection methods [22–24] are usually supervised. What is needed to train them are pairs consisting of a polarity score (e.g. *positive, neutral, negative* or scaled *1 to 5*) and text expressing the same polarity towards the entity. In addition, it is also useful to know the type of the described entity, e.g. its product category or type of local business, in order to learn specific models for different entity types.

Table 6: Distribution of schema:org/Review entities over different domains the WDC Microdata 2018 corpus

schema.org Type	Reviews		#Type entities
	#	%	
Product	6,371,735	46.96%	1,814,687
LocalBusiness	1,762,858	12.99%	455,200
Thing	1,405,100	10.35%	1,403,613
Organization	700,550	5.16%	233,216
Restaurant	465,208	3.43%	46,184
Hotel	368,831	2.72%	42,362
Book	350,826	2.59%	161,848
Service	217,817	1.61%	180,722
WebPage	144,441	1.06%	34,104
Place	60,348	0.44%	37,552
Article	59,943	0.44%	26,559
Event	59,916	0.44%	42,597
Person	57,740	0.43%	57,555
MobileApplication	41,610	0.31%	18,830
CreativeWork	4,072	0.03%	2,176

In sum around 130 thousand websites that are covered by the WebDataCommons 2018 Microdata corpus use the schema.org vocabulary to annotate reviews (see lower part of Table 1). Figure 1 shows an example of how a review about the tent is annotated in the HTML code of the web page. The schema.org term *reviewValue* is used to annotate the polarity score that is assigned to the tent. The term *bestRating* determines the rating scale and the term *reviewBody* annotates the free text review. The first *itemType* annotation determines the type of the reviewed entity, e.g. product. The WebDataCommons 2018 Microdata corpus contains 13.5 million *schema.org:Review* entities[18] that annotate review values and review bodies and can thus be used to train sentiment analysis methods. Table 6 shows the distribution of these reviews depending on the type of entity that is reviewed. We see that the corpus contains 6.3 million reviewValue/reviewBody pairs about 1.8 million different products, as well as 1.7 million reviewValue/reviewBody pairs judging 455 thousand local businesses.

[15] http://data.dws.informatik.uni-mannheim.de/LD4IE/

[16] http://microformats.org/wiki/hcard

[17] http://data.dws.informatik.uni-mannheim.de/LD4IE/2015/data/

[18] The review data can be downloaded from: http://webdatacommons.org/structureddata/2018-12/stats/schema_org_subsets.html

There exists a large body of research on sentiment analysis [21–24]. However, to the best of our knowledge none of the approaches exploits semantically annotated reviews from the Web as supervision. Commonly used sources of training data for sentiment analysis are tweets which are for instance used for the SemEval-2017 Task 4 [25]. The SemEval-2017 training sets consist of 20,000 to 50,000 text/polarity pairs depending on the specific subtask. A large collection of recommender systems datasets[19] has been collected by Julian McAuley. The datasets contain for instance reviews about products (e.g. 82.83 million reviews crawled from Amazon between 1996 and 2014), local businesses (e.g. 11.45 million reviews from Google maps) and books (1.5 million reviews from GoodReads, 2017). Compared to these datasets, using semantically annotated reviews from the Web as training data has the advantage that the reviews cover many languages [11,22], cover more entity types (e.g. also hotels, events, services), originate from a larger number or sources, and are more up-to-date.

9 Conclusion

This article has demonstrated the potential of using semantic annotations from the Web as training data for supervised matching methods. In addition, we also explored the potential of using semantic annotations as training data for information extraction and sentiment analysis. The experiments in Section 6 clearly showed the usefulness of the training data for the task of product matching despite of the dataset containing some noise (see error analysis in Section 4).

While the generated training dataset is already large, it has been built using only the tip of the iceberg as the Common Crawl only covers 3.1 billion HTML pages while commercial crawls are believed to cover at least one order of magnitude more pages. Thus, if specific experiments require more data, it is clearly possible to crawl deeper into the websites that we identified to annotate specific types of data and retrieve large quantities of additional data.

References

1. Achichi, M., Cheatham, M., Dragisic, z., et al.: Results of the ontology alignment evaluation initiative 2017. In: Proceedings of the 12th ISWC Workshop on Ontology Matching. pp. 61-113 (2017).
2. Suganthan, P., Doan, A., et al.: Falcon: Scaling Up Hands-Off Crowdsourced Entity Matching to Build Cloud Services. In: Proceedings of the 2017 ACM International Conference on Management of Data, pp. 1431-1446 (2017).
3. Daskalaki, E., Flouris, G., Fundulaki, I., Tzanina, S.: Instance matching benchmarks in the era of Linked Data. Journal of Web Semantics, 39, C, 1-14 (2016).
4. Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., Tang, N.: Distributed representations of tuples for entity resolution. Proceedings of the VLDB Endowment. 11, 11, 1454-1467 (2018).
5. Foley, J., Bendersky, M., Josifovski, V.: Learning to Extract Local Events from the Web. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 423-432 (2015).
6. Gokhale, C. et al.: Corleone: hands-off crowdsourcing for entity matching. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD 14. pp. 601-612 ACM Press, Snowbird, Utah, USA (2014).
7. Kärle, E., Fensel, A., Toma, I., Fensel, D.: Why Are There More Hotels in Tyrol than in Austria? Analyzing Schema.org Usage in the Hotel Domain. In: Proceedings of the International Conference on Information and Communication Technologies in Tourism 2016. pp. 99-112 Springer International Publishing (2016).
8. Konda, P. et al.: Magellan: toward building entity matching management systems over data science stacks. Proceedings of the VLDB Endowment. 9, 13, 1581-1584 (2016).
9. Köpcke, H., Thor, A., Rahm, E.: Evaluation of entity resolution approaches on real-world match problems. Proceedings of the VLDB Endowment. 3, 12, 484-493 (2010).
10. Das, S. et al.: The Magellan Data Repository, https://sites.google.com/site/anhaidgroup/useful-stuff/data.
11. Meusel et. al.: The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In: Proceedings of the International Semantic Web Conference. pp 277-292 (2014).
12. Meusel, R., Paulheim, H.: Heuristics for Fixing Common Errors in Deployed schema.org Microdata. In: The Semantic Web. Latest Advances and New Domains. pp. 152-168 Springer International Publishing (2015).
13. Meusel, R., Paulheim, H.: Creating Large-scale Training and Test Corpora for Extracting Structured Data from the Web. In: Proceedings of Third Workshop on Linked Data for Information Extraction (2015).
14. Mudgal, S. et al.: Deep Learning for Entity Matching: A Design Space Exploration. In: Proceedings of the 2018 International Conference on Management of Data - SIGMOD 18. pp. 19-34 ACM Press, Houston, TX, USA (2018).
15. Petrovski, P., Bizer, C.: Extracting attribute-value pairs from product specifications on the web. In: Proceedings of the International Conference on Web Intelligence - WI 17. pp. 558-565 ACM Press, Leipzig, Germany (2017).
16. Petrovski, P., Bryl, V., Bizer, C.: Integrating product data from websites offering microdata markup. In: Proceedings of the 23rd International Conference on World Wide Web - WWW 14 Companion. pp. 1299-1304 ACM Press, Seoul, Korea (2014).
17. Petrovski, P., Primpeli, A., Meusel, R., Bizer, C.: The WDC Gold Standards for Product Feature Extraction and Product Matching. In: Proceedings of the International Conference on E-Commerce and Web Technologies. pp. 73-86 Springer International Publishing, Cham (2017).
18. Qiu, D., Barbosa, L., Dong, X.L., Shen, Y., Srivastava, D.: Dexter: large-scale discovery and extraction of product specifications on the web. Proceedings of the VLDB Endowment. 8, 13, 2194-2205 (2015).

[19] https://cseweb.ucsd.edu/~jmcauley/datasets.html

19. Ristoski, P., Petrovski, P., Mika, P., Paulheim, H.: A machine learning approach for product matching and categorization. Semantic Web. 9, 5, 707-728 (2018).
20. Shah, K., Kopru, S., Ruvini, J.D.: Neural Network based Extreme Classification and Similarity Models for Product Matching. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers). pp. 8-15 Association for Computational Linguistics, New Orleans - Louisiana (2018).
21. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. 5, 1, 1-167 (2012).
22. Deriu, J. et al.: Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In: Proceedings of the 26th International Conference on World Wide Web - WWW 17. pp. 1045-1052 ACM Press, Perth, Australia (2017).
23. Severyn, A., Moschitti, A.: Twitter Sentiment Analysis with Deep Convolutional Neural Networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 15. pp. 959-962 ACM Press, Santiago, Chile (2015).
24. Tang, D. et al.: Sentiment Embeddings with Applications to Sentiment Analysis. IEEE Transactions on Knowledge and Data Engineering. 28, 2, 496-509 (2016).
25. Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 task 4: Sentiment analysis in Twitter. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 502-518, 2017.