Towards Multi-modal Entity Resolution for Product Matching

Moritz Wilke
Leipzig University
wilke@informatik.uni-leipzig.de

Erhard Rahm
Leipzig University
rahm@informatik.uni-leipzig.de

ABSTRACT

Entity Resolution has been applied successfully to match product offers from different web shops. Unfortunately, in certain domains the (textual or numerical) attributes of a product are not sufficient for a reliable match decision. To overcome this problem we extend an attribute-based matching system to incorporate image data, which are available in almost every web shop. To evaluate the system we enhance the WDC product matching dataset with images crawled from the web. First evaluations show that the use of images is beneficial to increase recall and overall match quality.

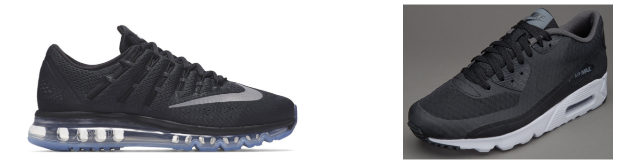
Keywords

Record Linkage, Product Matching, Deep Learning

1. INTRODUCTION

Entity resolution (ER), also known as record linkage, is the procedure of identifying which items from one or more data sources refer to the same real-world entities. It is an important step in data integration where the goal is to unify data from different origins to increase the quality and size of available data for further analysis. It is mostly based on structured or semi-structured entity descriptions consisting of several attributes (such as 'Name', 'Date of Birth, 'Zip code'). For web data the attributes are often missing and noisy and may contain longer textual descriptions.

An application of ER is the matching of product offers across the web. It can be used to compare prices and inventories of web shops or to present the best offer for a desired product to a user. This task can often be tackled with attribute-based ER, e.g. utilizing attribute values extracted from the product web pages. A smart phone for example can be distinguished by attributes such as 'brand', 'model', 'storage size', 'display size', 'weight', etc. But it is still hard to match items where (textual) information is sparse or varies much, e.g. in the fashion domain. In contrast, the selling of fashion items is largely driven by their visual features and hence every shop has images available.



Title	nike air max 2016 806771 001 para hombre negro 040 footstop	nike sportswear air max 90 ultra essential black mens shoes dark grey white 819474 013
Desc.	<None>	featuring no sew overlays the air max 90 ultra delivers a supportive and lightweight feel its visible air sole unit helps absorb impact […]
Brand	<None>	<None>
Price	<None>	<None>

Figure 1: Example of two (non-matching) products

Recent developments in computer vision (which are largely driven by deep learning technologies) allow the assumption that a reliable distinction of product images is achievable, given enough data. The two similar but different shoes presented in Figure 1 highlight some problems and opportunities of visual product matching. The images readily show that the two shoes are different while the attribute information makes it difficult to come to a clear decision. This is because the missing attribute values lead to larger differences in the description. Moreover, the description of the second shoe is written in an advertising way to convince the customer, which decreases its informational value. So in this case the title attribute is the only usable attribute which does not contain a lot of information, however. This leads to our main research question: *How can we utilize the additional information provided by product images to improve matching quality?*

An obstacle for research in ER in general and product matching has been the lack of large and public datasets that contain ground-truth information about matching entities. There are several datasets that contain product images in combination with descriptions and there are datasets that contain matching pairs of either (product) images or descriptions. However, there is currently no public dataset that contains product images and descriptions as well as the true set of matching items.

The main contributions of this work are the creation of a suitable multi-modal ER benchmark dataset (Section 3), an extension of the existing DeepMatcher [11] ER framework to also use image data for matching (Section 4), and a first evaluation of the system on a subset of the dataset (Section 5). As this work is still in progress we also discuss current

32^{nd} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 25.05.2021 - 28.05.2021, Grimma, Germany.

limitations and plans for future work (Section 6).

2. RELATED WORK

This work builds on previous work on ER approaches, existing ER datasets, and neural networks for text and image analysis.

2.1 Entity Resolution

There is a long history of research in the field of entity resolution and a comprehensive introduction can be found in [4]. A lot of work has been put into adequate similarity metrics for attribute values and rule- or tree-based combinations of those similarities for match classification. Furthermore, blocking techniques have been devised to pre-filter match candidates in order to reduce the quadratic complexity of comparing all items with each other.

In the last years ER has been scaled to large datasets using the map-reduce paradigm and there has been a lot of research on utilizing machine learning for either parts of an ER pipeline (e.g. blocking, similarity computation, match decisions) or the configuration of the whole process. Due to the increasing amount of data available from the internet, ER approaches that work on heterogeneous, noisy and unstructured or semi-structured data have gained importance [12].

An overview of the (increasing) usage of neural networks and deep learning for entity resolution can be found in [2]. A recent ER system that uses deep learning is DeepMatcher [11]. It can use word-embeddings to encode the information from different attributes and provides different sequence models to align and compare those encodings.

Applying ER on e-commerce data has been the focus of Köpcke et al. [8]. An initial approach to use image data for product matching can be found in [14]. The authors use attribute matching and enrich it with image embeddings generated from a convolutional neural network. The main difference to our approach is that the neural network is not directly trained in the matching task but solely functions as a feature extractor. The cosine similarity between two image embeddings is passed to a match classifier along with the similarities of textual features. An evaluation is performed on three datasets from the electronics domain (laptops, televisions, phones) which consist of 200 - 300 products. F-scores of up to 73.35% (laptops), 83.27% (phones) and 84.96% (televisions) are achieved by using the images to improve the best text matching results by a small amount (around 1%). The authors conclude that images cannot be used as a strong matching signal because some product variations that are important to distinguish (such as a phone with 16GB of storage vs. one with 64GB) use the exact same image. Such problems should have less impact in other product domains such as clothing. Current research in computer vision and image retrieval allows to expect a robust image matching given enough clean data at least for certain categories of products.

2.2 Datasets

The WDC dataset (WDC Product Data Corpus and Gold Standard for Large-scale Product Matching 2.0) has been published by Primpeli et al. [13]. They use the (partial) existence of product identifiers such as MPN and SKU to create weak clusters of matching products. To refine these noisy clusters and to obtain training data and ground-truth for evaluation they apply a number of heuristics, machine learning algorithms and manual processing steps. The result is a dataset of 16 million products (supposedly described in English). The applied clustering by identifiers can be considered as a silver standard. The creators add handcrafted true matches and semi-automatically created differently sized training sets in four product categories (watches, shoes, cameras, electronics). Although these datasets are suitable for ER evaluations, they do not contain product images. The aforementioned experiments by Ristoski et al. [14] have been conducted on an earlier version of the WDC dataset, but unfortunately the corresponding images are not available anymore.

DeepFashion2 [5] is a dataset for image retrieval in the fashion domain. It contains different images for the same product, some from shops some from users. However, the products have no textual properties (such as a description, brand or price) so that it is not multi-modal.

The dataset for the SIGIR eCom 2020 multi-modal product classification and retrieval challenge [1] contains product descriptions and images but does not contain ground truth for product matching.

To the best of our knowledge, there is currently no ER dataset with both product images and descriptions *and information about matching products*.

2.3 Multi-modal deep learning

Over the last years, convolutional neural networks (CNN) have achieved many breakthroughs in the field of computer vision and have contributed much to the wide adoption of deep learning [9]. An adaption of deep learning for fashion images called Match-R-CNN is presented in [5]. It supports the identification and classification of fashion items as well as image retrieval. For attribute-based ER, deep learning approaches internally work mostly with the concept of distributed representations (embeddings). The combination of text and image data in deep learning systems is called multimodal deep learning. Typical applications include the creation of text description for images or image retrieval from text queries. An overview on tasks, datasets and problems in this field can be found in [10].

3. BENCHMARK DATASET CREATION

To overcome the mentioned lack of a multi-modal ER benchmark dataset, we extend the WDC dataset with image data. The underlying common crawl[1] snapshot dates back to November 2017 and does not contain additional data apart from the HTML documents, hence it is not initially clear to which amount the URLs are still valid and whether the images are still available. To retrieve images, the following procedure is used: First the documents are parsed for HTML tags that contain image URLs and a (semantic) annotation that indicates that the images belongs to a product (this is needed to avoid collection unrelated images from the website).

In the second step, we query the URLs to retrieve up to five images per product. Finally a procedure to query the internet archive[2] for missing images is used. Due to its slow speed this method is only applied selectively to achieve higher coverage of images for the experiments.

[1] http://commoncrawl.org/

[2] https://archive.org/

	products	**positive pairs**	**negative**	**% title**	**% description**	**% image**
WDC shoes train (xlarge)	2450	4141	38288	100%	53%	79%
WDC shoes gold standard	1111	300	800	100%	69%	81%
Evaluation subset train	950	1350	6286	100%	52%	100%
Evaluation subset gold standard	813	206	586	100%	76%	100%

Table 1: Comparison of the original WDC shoe training and gold standard data and the cleaned evaluation subset for which all images have been verified manually.

The results is a *database of images for 10M (63%) products from the WDC corpus.* The gathered image data is far from clean. Problems include that some images do not depict the correct product but a (brand or shop) logo, a placeholder or something entirely different, or that images show the product isolated or in context of a scene and in combination with other objects. Also some images are pixel-wise duplicates of each other.

Table 1 shows the crawling results for the shoe category of the original dataset. The initial coverage of raw images is 79% on the training set and 81% on the gold standard. Since we want to run our initial evaluation on data that has full image coverage and does not contain wrong or noisy image data, we manually verify the images and create evaluation subsets with the desired properties which are subsequently smaller. The inspection also reveals that pairs in the existing gold standards have been determined without consideration of the product images. Hence, there are hand-labelled match pairs with a similar textual description where the images show (minor) differences. A similar problem occurs in the training data: to create training pairs with optimal informational value, the authors of the original dataset adapt a method from [7]. It consists of the precomputation of similarity scores between known matches and non-matches, which is followed by choosing positive training pairs with low and negative training pairs with high similarity. Since we do not repeat this process, the images have no influence on the choice of the training pairs. The effect of this bias has not been examined yet.

4. MATCHING SYSTEM

Our current system for multi-modal product matching builds on DeepMatcher [11]. We add the capability of processing image data while keeping the overall structure and components. Figure 2 shows the match processing for a pair of records with the same attributes and an image to determine a binary match/non-match decision. Schema alignment and the creation of suitable candidate pairs are not part of the system.

We now describe the four main components of DeepMatcher and how they are related to the processing of images. The first component is *Attribute-level embedding*, which converts each word (or n-gram) of an attribute value into a word vector by applying a pre-trained word model. The output of this component is a list of such embeddings with the same length as the the number of input words. Because those lists are of different length for each record, they have to be aligned before the attributes of two products can be compared. This procedure is performed by the *attribute summarizer*, which can be any kind of sequence to vector module, e.g. a recurrent neural network. The objective of this module is to compress the information, filter out redundant and meaningless words and represent the attribute as a fixed-length vector.

An image processing module is added to DeepMatcher with the aim to create an image representation that can be processed exactly like the existing feature vectors. This is achieved in the following way: first an optional pre-processing step detects the main shapes in the image and crops it accordingly. This is done to reduce white-space and and non-informational regions of the image. There are other possible ways to achieve this cropping, for example by using another neural network for object detection and/or masking the relevant regions in the image. But these methods add further complexity and they need training data (bounding boxes or masks) for each product category. The second step is to feed the image to a pre-trained image classification neural network (our first experiments use Resnet50 [6]) and to append a fully connected layer to downsample its representation of the image to the dimension of the other feature vectors. To reduce the number of learnable parameters and hence training time, the first six layers of the backbone neural network are fixed and do not change during the training of the matching system. Since we currently only use a single image per product, summarization is not needed for the images.

After these steps, the feature vector resulting from an image or a text attribute has the same dimensions and can be treated equally. The *attribute comparators* take feature vectors of the same attribute and from both products as input and create a similarity representation. In DeepMatcher, this can be the absolute distance of both vectors, their concatenation or any other method that returns a vector. All attribute similarities are finally fed into the classification component which returns the final match decision. The standard classification method is a two-layered, fully connected neural network.

5. EVALUATION

To evaluate the system, a subset of the existing WDC shoe training set and gold standard is created by filtering all pairs that can be enriched with image data for both products. The size and attribute coverage of the evaluation datasets are provided in Table 1. This may not reflect the real-world data, but it is easier to establish a working system on clean and complete data and later improve its robustness against real-world anomalies. Similarly, the shoe category was chosen as initial category because we assume that products in a domain such as fashion that is determined by visual factors are more likely to match using their images than products from a category like electronics where the visual aspect plays a minor role and the products are easier to describe with technical specifications.

Selecting the pairs with images, decreases the size of the training set from 42.4k pairs to 7.6k pairs and the size of the

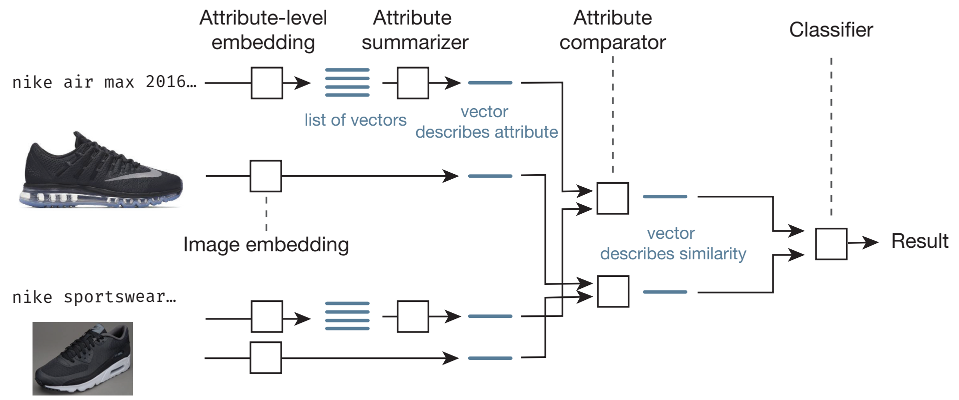


Figure 2: Workflow for multi-modal product matching

Features	F1	P	R
image	73.1	61.2	90.8
title	85.2	79.3	91.9
title+image	85.6	79.2	93.0
title+description	81.5	73.2	91.9
title+description+image	83.6	75.4	93.8

Table 2: Match quality on the (clean) shoe dataset

gold standard (used as test set) from 1,111 to 792 pairs. A validation set is taken out of the training data.

We choose the DeepMatcher configuration that performed best on the baseline experiments for the WDC dataset conducted in [13], which is using a pre-trained fastText model ([3]) to create token-level embeddings from the input data and the RNN method for attribute summarization. Different feature combinations of title, description and image values are tested. The model is trained for 20 epochs on a GeForce RTX 2080 Ti GPU. When using only the product images for matching we observe that the system needs a longer time to converge, hence we use 40 epochs of training in that case. Each feature combination is trained and evaluated three times and the presented scores are the mean results of these runs.

The experiments in [13] show that DeepMatcher outperforms methods such as support vector machines, logistic regression and random forest, which are also used in [13]. Hence we do not compare directly against these methods.

The matching quality is measured as precision, recall and f-score, as presented in Table 2. Using only the images achieves a recall of over 90% and an f-score of 73% which shows that the images provide valuable information for matching. Using the textual attributes alone is more effective and the use of attribute title performs better than using both title and description. This is influenced by the fact that the description is noisy and missing in many cases so that it is of limited usefulness. Using the images in addition to the textual attributes improves recall by up to about 2% and also the f-score. The best overall f-score of 85.6% is achieved by combining title and image similarity albeit the use of images allowed for only a small improvement here (0.4%). For the use of title and description, the additional use of images enabled a bigger f-score improvement by 2.1%.

While these improvements appear modest one has to bear in mind that the underlying product dataset has been created for attribute-based ER and can already achieve high match quality for the use of attribute values only. Furthermore, as mentioned in Section 3, the images have not been present at the time of labelling, hence the golden truth is still not fully consistent with regards to images. This is also illustrated in Figure 1 where the images clearly show different shoes while the mismatch is harder to infer from the title attributes.

6. FUTURE WORK

Concerning the image crawling, our goal is to increase coverage of images and manually review and relabel the existing datasets to create and publish a dataset that can be used for further experiments and evaluation. The resulting dataset is likely of high value not only for research on data cleaning, product matching and retrieval but also for other applications such as product classification and multi-modal tasks e.g. the generation of product descriptions from images.

The matching system can be improved by different preprocessing steps and matching architectures as well as the use of multiple images per product. Our current system (as most ER systems) compares entities attribute-wise and aggregates attribute similarities for a match decision. This follows from the reasoning in structured ER that every attribute describes a single feature of the record that can only be compared with the corresponding field of another object (e.g. if a company called ”blue” offers a yellow product, it is not similar to a blue product of a different company). But this assumption does not hold for the different features in our case: a product description text and a product image can be interpreted as different encodings of the same information. The equal treatment of text and image features as embeddings in our system allows to compare *across the modalities*, which might be a good method to overcome sparsity of images and descriptions and lead to better matches.

7. CONCLUSION

We show that the use of images is a promising strategy to improve product matching results in the fashion domain. Although the improvements in f-score are only minor, the image data can be used to obtain a good recall of matching pairs. Further experiments and data preparation have to be conducted to validate this hypothesis and ensure the quality of the matching and its stability under real-world conditions.

A new multi-modal product matching dataset that is created for this experiment can be of use for researchers in many applications.

8. ACKNOWLEDGMENTS

This project was funded by the Sächsische Aufbaubank (FKZ 100378106) and German BMBF within the Project ScaDS.AI Dresden/Leipzig (BMBF 01IS18026B). Computations for this work were done (in part) using resources of the Leipzig University Computing Centre.

9. REFERENCES

[1] H. Amoualian, P. Goswami, L. Ach, P. Das, and P. Montalvo. SIGIR 2020 E-Commerce Workshop Data Challenge. 2020.

[2] N. Barlaug and J. A. Gulla. Neural Networks for Entity Matching. 2020.

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *CoRR*, 2016.

[4] P. Christen. *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer Berlin Heidelberg, 2012.

[5] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo. DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[7] H. Köpcke and E. Rahm. Training selection for tuning entity matching. *Proceedings of the Sixth International Workshop on Quality in Databases and Management of Uncertain Data (QDB/MUD)*, 2008.

[8] H. Köpcke, A. Thor, and E. Rahm. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 2010.

[9] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521, 05 2015.

[10] A. Mogadala, M. Kalimuthu, and D. Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv*, 2019.

[11] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra. Deep learning for entity matching: A design space exploration. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2018.

[12] G. Papadakis, E. Ioannou, and T. Palpanas. Entity Resolution: Past, Present and Yet-to-Come From Structured to Heterogeneous, to Crowd-sourced, to Deep Learned. In *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT)*, 2020.

[13] A. Primpeli, R. Peeters, and C. Bizer. The WDC training dataset and gold standard for large-scale product matching. In *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 2019.

[14] P. Ristoski, P. Petrovski, P. Mika, and H. Paulheim. A machine learning approach for product matching and categorization. *Semantic Web*, 9, 2018.