Automated Extraction of Fine-Grained Standardized Product Information from Unstructured Multilingual Web Data

Alexander Flick[1(✉)], Sebastian Jäger[1], Ivana Trajanovska[1], and Felix Biessmann[1,2]

[1] Berlin University of Applied Sciences and Technology, Berlin, Germany
alexander.flick@bht-berlin.de
[2] Einstein Center Digital Future, Berlin, Germany

Abstract. Extracting structured information from unstructured data is one of the key challenges in modern information retrieval applications, including e-commerce. Here, we demonstrate how recent advances in machine learning, combined with a recently published multilingual data set with standardized fine-grained product category information, enable robust product attribute extraction in challenging transfer learning settings. Our models can reliably predict product attributes across online shops, languages, or both. Furthermore, we show that our models can be used to match product taxonomies between online retailers.

Keywords: Product information extraction · E-commerce

1 Introduction

Recent research achievements in the field of machine learning (ML) [1,13] have the potential to improve automated information extraction in applications such as e-commerce. However, the translation of these ML innovations into real-world application scenarios is impeded by the lack of publicly available data sets. Here we demonstrate that recent advances in ML can be translated into automated information extraction applications when leveraging carefully curated data. To better assess the contribution of this study, we first highlight some relevant data sets and methods that aim at the automated extraction of structured data in the field of e-commerce.

Public E-commerce Data Sets. We summarize publicly e-commerce data sets used for the automated extraction of product information in Table 1. To leverage the potential of ML, large and diverse data sets that follow a fine-grained product taxonomy are favorable. A common and detailed taxonomy is the Global Product Classification (GPC) standard, which "classifies products by grouping them into categories based on their essential properties as well as their relationships to

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13982, pp. 230–235, 2023.
https://doi.org/10.1007/978-3-031-28241-6_19

Table 1. Comparison of e-commerce data sets used for product attribute extraction and classification. Column *GPC* means whether or not the data set follows the GPC taxonomy.

	Regular	Multi-			GPC	Size
	Updated	Lingual	Shop	Family		
Farfetch product meta data [9]	✗	✗	✗	✗	✗	400 K
Product details on Flipkart [3]	✗	✗	✗	✓	✗	20 K
Amazon browse node classification [2]	✗	✗	✗	✓	✗	3 M
Amazon product-question answering [16]	✗	✗	✗	✓	✗	17.3 GB
Rakuten data challenge [10]	✗	✗	✗	✓	✗	1 M
MAVE [18]	✗	✗	✗	✓	✗	2.2 M
Innerwear from victoria's secret & co [15]	✗	✗	✓	✗	✗	600 K
WDC-MWPD [19]	✗	✗	✓	✗	✓	16 K
WDC-25 gold standard [14]	✗	✗	✓	✓	✓	24 K
GreenDB [7]	✓	✓	✓	✓	✓	>576 K

other products" [4]. For example, multiple *Brick*s (shirts and shorts) can belong to the same *Family* (clothing) but are different *Class*es (upper and lower body wear)[1].

Multilingual Fine-Grained Product Classification. There are few recent studies investigating automated extraction of standardized product information in text corpora. Brinkmann et al. [1] study how hierarchical product classification benefits from domain-specific language modeling. They report an improvement of 0.012 weighted F1 score by using schema.org product[2] annotations for pre-training. Peeters et al. [12] study cross-language learning for entity matching and demonstrate that multilingual transformers outperform single-language models (German BERT) by 0.143 F1 when trained on a single language (German) and tested on multiple (German and English). Furthermore, using additional training data for the second language (English) improves the performance by another 0.038 weighted F1.

These studies highlight the potential of modern ML methods for automated product attribute extraction. In this work, we show that transfer learning helps to extract structured information (product category) from unstructured data (product name and description) and to find reliable taxonomy mappings.

2 Experiments

We evaluate three transfer learning scenarios for product classification:

1. **Language Transfer:** training on data of one language, test on other language data.

[1] See the GPC Browser for more examples: https://gpc-browser.gs1.org/.
[2] Website: https://schema.org/Product.

2. **Shop Transfer:** training on data of one shop, test on other shop data.
3. **Language and Shop Transfer:** training on data of one shop and one language, test on data of different shops and languages.

Furthermore, we study whether ML methods can be used to find reliable taxonomy mappings. For this, we apply a model trained for a *target taxonomy* to data that uses a *source taxonomy*. For each source category, the majority of predicted target categories define the mapping from source to target taxonomy.

Data Sets. In our experiments, we use two data sets, the GreenDB [6] and the Farfetch data set [9]. The GreenDB[3] is a multilingual data set covering 5 European shops with about 576k unique products of the 37 most important product categories following the GPC taxonomy. It covers categories from the GPC segments Clothing, Footwear, Personal Accessories, Home Appliances, Audio Visual/Photography, and Computing. A recent publication [8] presents the GreenDB's high quality and usefulness for information extraction tasks. The Farfetch data set has about 400k unique products from a single shop. It does not follow a public taxonomy and covers only fashion products.

ML Model. The experiment implementation is based on autogluon's [17] TextPredictor and uses *mDeBERTav3* [5] as the backbone model. For training, we use the GreenDB and apply Cleanlab [11] to find and remove miss-classified products (211 were found). Our models use the product's name and description to predict their product category. $model_{baseline}$ is trained on the entire GreenDB (all shops), $model_{ZaDE}$ on the German, $model_{ZaFR}$ on the French, and $model_{ZaALL}$ on the German, French, and English Zalando products contained in the GreenDB.

Online Demo. To demonstrate the transfer capabilities, we published an online demo available: https://product-classification.demo.calgo-lab.de. As shown in Fig. 1, it automatically downloads the HTML of a given URL, extracts the products' name and description, and uses $model_{baseline}$ to predict its GPC category.

3 Results

The baseline performance ($model_{baseline}$) shows a strong 0.99 weighted F1 score on a GreenDB test set.

Transfer Tasks. $model_{ZaDE}$ demonstrates language transfer when it is applied to other languages of the same shop. It achieves weighted F1 scores of 0.898 for English and 0.873 for French. Applying $model_{ZaFR}$ and $model_{ZaDE}$ on other shops demonstrates shop transfer with weighted F1 scores from 0.648 to 0.836. If the model is fine-tuned on multi-lingual data ($model_{ZaALL}$), almost all shops benefit, see Table 2 for details. The language and shop transfer is even more challenging and performs worse for all shops. Transferring across data sets, i.e., applying $model_{baseline}$ to Farfetch data, achieves a 0.924 weighted F1 score.

[3] We use GreenDB version 0.2.2 available at https://zenodo.org/record/7225336.

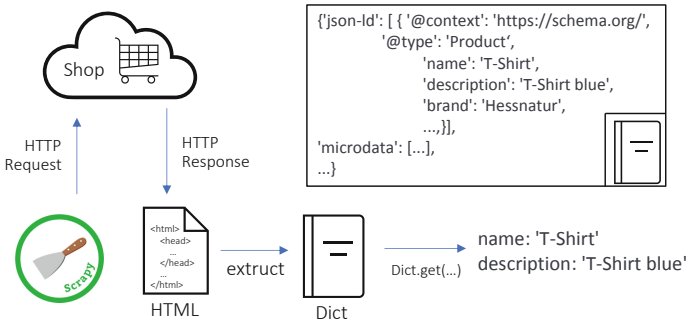

Fig. 1. Online demo overview. Automated extraction of schema.org information (product name and description) from HTML, used for product classification.

Table 2. Weighted F1 scores for shop transfer experiments. Scores from 0.648 to 0.836 demonstrate robust shop transfer. Shop transfer profits from additional data in other languages.

	Model	FR		DE	
		Asos	H&M	Otto	Amazon
Shop Transfer	$model_{ZaFR}$	0.836	0.678	–	–
	$model_{ZaDE}$	–	–	0.777	0.648
	$model_{ZaALL}$	0.842	0.717	0.762	0.739
Shop & Language Transfer	$model_{ZaFR}$	–	–	0.614	0.449
	$model_{ZaDE}$	0.795	0.666	–	–

Taxonomy Matching. Using $model_{baseline}$ to map products' categories from Farfetch to GreenDB (GPC taxonomy) results in 41 out of 46 (>89%) correctly mapped categories.

4 Conclusion

We demonstrate that combining rich multilingual data sets and modern ML methods enables fine-grained standardized product information extraction from unstructured data. We investigate several transfer learning settings when training and testing on data from different shops and languages, even in zero-shot scenarios when no data from another shop and language was available in the training data.

Acknowledgements. This research was supported by the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety based on a decision of the German Bundestag.

References

1. Brinkmann, A., Bizer, C.: Improving Hierarchical Product Classification using Domain-specific Language Modelling. IEEE Data Eng. Bull. **44**(2), 14–25 (2021). http://sites.computer.org/debull/A21june/p14.pdf
2. Challenge, A.M.: (2022). https://www.hackerearth.com/en-us/challenges/competitive/amazon-ml-challenge/, [Online; Accessed 23 May 2022]
3. Flipkart: (2022). https://www.kaggle.com/PromptCloudHQ/flipkart-products, [Online; Accessed 23 May 2022]
4. GS1: Global Product Classification (GPC) — GS1. https://www.gs1.org/standards/gpc, [Online; Accessed Oct 20 2022]
5. He, P., Gao, J., Chen, W.: Debertav 3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. CoRR abs/2111.09543 (2021). https://doi.org/10.48550/arxiv.2111.09543
6. Jäger, S., Greene, J., Jakob, M., Korenke, R., Santarius, T., Biessmann, F.: GreenDB: Toward a Product-by-Product Sustainability Database. Tech. rep., arXiv (May 2022). https://doi.org/10.48550/arXiv.2205.02908
7. Jäger, S., Bießmann, F., Flick, A., Sanchez Garcia, J.A., von den Driesch, K., Brendel, K.: GreenDB: A Product-by-Product Sustainability Database (Feb 2022). https://doi.org/10.5281/zenodo.6576662,Supported by the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety based on a decision of the German Bundestag. Förderkennzeichen: 67KI2022B
8. Jäger, S., Flick, A., Garcia, J.A.S., Driesch, K.v.d., Brendel, K., Biessmann, F.: GreenDB - A Dataset and Benchmark for Extraction of Sustainability Information of Consumer Goods (Aug 2022). https://doi.org/10.48550/arXiv.2207.10733
9. Kale, A., Kallumadi, S., King, T.H., Malmasi, S., de Rijke, M., Tagliabue, J.: Ecom'22: The sigir 2022 workshop on ecommerce. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3485–3487. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3531701
10. Lin, Y., Das, P., Datta, A.: Overview of the SIGIR 2018 ecom rakuten data challenge. In: Degenhardt, J., Fabbrizio, G.D., Kallumadi, S., Kumar, M., Trotman, A., Lin, Y., Zhao, H. (eds.) The SIGIR 2018 Workshop On eCommerce co-located with the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018), Ann Arbor, Michigan, USA, July 12, 2018. CEUR Workshop Proceedings, vol. 2319. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2319/ecom18DC_paper_13.pdf
11. Northcutt, C.G., Jiang, L., Chuang, I.L.: Confident learning: estimating uncertainty in dataset labels. J. Artif. Intell. Res. **70**, 1373–1411 (2021). https://doi.org/10.1613/jair.1.12125
12. Peeters, R., Bizer, C.: Cross-Language Learning for Entity Matching. In: Companion Proceedings of the Web Conference 2022, pp. 236–238 (Apr 2022). https://doi.org/10.1145/3487553.3524234
13. Peeters, R., Bizer, C., Glavas, G.: Intermediate training of BERT for product matching. In: Piai, F., Firmani, D., Crescenzi, V., Angelis, A.D., Dong, X.L., Mazzei, M., Merialdo, P., Srivastava, D. (eds.) Proceedings of the 2nd International Workshop on Challenges and Experiences from Data Integration to Knowledge Graphs co-located with 46th International Conference on Very Large Data Bases, DI2KG@VLDB 2020, Tokyo, Japan, August 31, 2020. CEUR Workshop Proceedings, vol. 2726. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2726/paper1.pdf

14. Primpeli, A., Peeters, R., Bizer, C.: The wdc training dataset and gold standard for large-scale product matching. In: Companion Proceedings of The 2019 World Wide Web Conference, pp. 381–386. WWW '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3308560.3316609
15. PromptCloud: Innerwear Data from Victoria's Secret and Others. https://www.kaggle.com/datasets/PromptCloudHQ/innerwear-data-from-victorias-secret-and-others (2022). [Online; Accessed Oct. 20 2022]
16. Rozen, O., Carmel, D., Mejer, A., Mirkis, V., Ziser, Y.: Answering product-questions by utilizing questions from other contextually similar products. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 242–253. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.23
17. Shi, X., Mueller, J., Erickson, N., Li, M., Smola, A.J.: Benchmarking multimodal automl for tabular data with text fields. CoRR abs/2111.02705 (2021). https://doi.org/10.48550/arxiv.2111.02705
18. Yang, L., et al.: MAVE: A Product Dataset for Multi-source Attribute Value Extraction. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 1256–1265. WSDM '22, Association for Computing Machinery, New York, NY, USA (Feb 2022). https://doi.org/10.1145/3488560.3498377
19. Zhang, Z., Bizer, C., Peeters, R., Primpeli, A.: MWPD2020: semantic web challenge on mining the web of html-embedded product data. In: Zhang, Z., Bizer, C. (eds.) Proceedings of the Semantic Web Challenge on Mining the Web of HTML-embedded Product Data co-located with the 19th International Semantic Web Conference (ISWC 2020), Athens, Greece, November 5, 2020. CEUR Workshop Proceedings, vol. 2720. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2720/paper1.pdf