GAVI: A Category-Aware Generative Approach for Brand Value Identification

Kassem Sabeh
Free University of Bozen-Bolzano
ksabeh@unibz.it

Mouna Kacimi
Wonder Technology Srl
mouna@wonderflow.ai

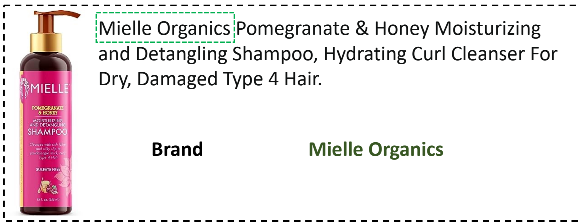
Johann Gamper
Free University of Bozen-Bolzano
jgamper@unibz.it

Abstract

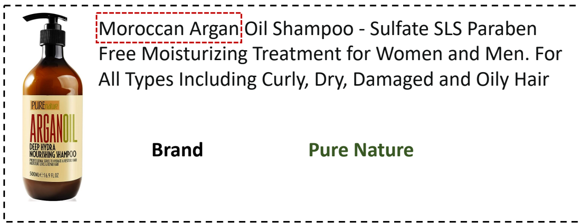
Extracting product attribute value information is vital for many e-commerce applications. One of the most crucial product attributes is the brand, as it significantly impacts customers' purchasing decisions and behaviour. Consequently, it is critical for e-commerce platforms to automatically and accurately identify brand values from product descriptions. Most existing methods focus on brand value extraction from text descriptions using sequence tagging and question answering techniques. However, brand values are often not mentioned explicitly in the product descriptions. Also, these approaches are designed without paying attention to product categories, which are important for brand value identification. In this work, we propose a novel category-aware generative approach for brand value identification (GAVI). In particular, we formulate the brand value identification problem as a sequence-to-sequence generation task. We use the T5 language model as the backbone of our approach. This allows us to identify brand values that are not explicitly mentioned in the title in a generative manner. We then propose to highlight the product categories inside our model input, making the approach category-aware. We conduct extensive experiments on a public dataset for brand value identification. The experimental results demonstrate that our generation-based approach outperforms existing extraction-based methods. Our code is released along with the fine-tuned models presented in the paper[1], which are also available as a demo[2].

1 Introduction

Product attributes are a crucial component of e-commerce platforms as they provide valuable information for customers to browse and compare products. One of the most important product attributes is the brand, as it plays a pivotal role in influencing customers' behaviour and purchasing decisions (Chovanová et al., 2015; Shahzad et al., 2014). Brand names also increase the recognisability of products and services amongst consumers, and permit them to deduce knowledge about important features of the product (Zhang et al., 2015). For instance, Figure 1a shows an example of a profile of a shampoo product taken from an e-commerce website. The brand of this product is "Mielle Organics". Knowledge about the brand can help the customers build a set of associations, like this shampoo is of "*high quality natural and organic ingredients*", is tailored for "*frizzy or curly hair*", has a "*Moisturizing and Detangling effect*" and is designed "*for women*". Consequently, when customers shop for a shampoo, they often select a particular brand based on their prioritized attributes and features. These inherent correlations between brands and product attributes underscore the critical need for e-commerce applications to automatically and accurately identify brand names from product descriptions.



(a) Brand value is mentioned in the product title.



(b) Brand value can not be extracted from the product title.

Figure 1: Examples of brand values in two product profiles.

[1] https://github.com/kassemsabeh/gavi
[2] https://bit.ly/3FHZGjU

Existing work for brand value identification falls under the general problem of attribute value extraction from product titles, with a plethora of research being developed to tackle this problem (Putthividhya and Hu, 2011; Kozareva et al., 2016; Zheng et al., 2018; Xu et al., 2019; Wang et al., 2020). Early approaches for attribute value extraction rely on rule based techniques and domain-specific dictionaries (Ghani et al., 2006; Vandic et al., 2012; Kozareva et al., 2016). These methods carry a close-world assumption and do not work well with new values, since they need to develop rules for every possible value. Consequently, they are not suitable for brand value identification where new brands are constantly emerging. With the advent of natural language understanding, sequence labeling methods have been developed (Huang et al., 2015; Zheng et al., 2018; Sabeh et al., 2022b). These methods utilize a BiLSTM-CRF architecture similar to NER tasks. However, their performance on attribute value extraction is limited by the abundance of negative token labels (e.g., the 'O' in BIO schema), which leads to many false negative results. Recently, question answering (Xu et al., 2019; Wang et al., 2020; Yang et al., 2022) based approaches were proposed. These methods scale existing sequence based methods to deal with multiple attribute inputs. All of the above approaches achieve promising results, however, they suffer from two major limitations:

- Most of the existing approaches are extractive-based methods; i.e., they extract the brand values from the text descriptions in the product profile. However, target brand values are sometimes absent from the textual descriptions of the product. For example, in Figure 1b, the brand of the product is "Pure Nature", which is not explicitly mentioned in the product title. The existing models extract instead the value "Moroccan Argan" as the brand, which is the wrong value.

- Existing methods for brand value extraction are designed without considering the product category. This is crucial for determining the set of applicable values, because categories can be substantially different in terms of brand values. For example, the value "Sunflower" can be a brand in the *Clothing* category. However, it used to indicate the scent in the *Food* category.

In this paper, we formulate the brand value identification problem as a sequence-to-sequence generation task. Inspired by the recent advances on text generation (Ushio et al., 2022; Bao et al., 2020; Xiao et al., 2021), we propose a category-aware **G**enerative **A**pproach for brand **V**alue **I**dentification, namely GAVI. In contrast to previous extractive approaches, we employ T5 (Raffel et al., 2020) language model as the backbone of our sequence-to-sequence approach. This generative approach allows the extraction output to expand beyond strings and sub-strings mentioned in the product textual description, which addresses the first limitation. To make the model category-aware, we propose to highlight the product categories inside the model input. This setup fits naturally with the sequence-to-sequence model architecture, and allows us to learn category-specific token embeddings that are effective for our task. We summarize the main contributions of our work as follows:

- We propose GAVI, a generative sequence-to-sequence model to identify brand values from product descriptions. To the best of our knowledge, this is the first work for generative brand value identification.

- We extend the basic generative solution to a category-aware sequence-to-sequence model by highlighting the product categories inside the input.

- We conduct extensive experiments on a public dataset, demonstrating the effectiveness of the proposed approach over several state-of-the-art baselines.

2 Related Work

Early work on attribute value extraction relied on rule-based techniques (Nadeau and Sekine, 2007; Vandic et al., 2012; Gopalakrishnan et al., 2012), which utilize domain-specific seed dictionaries to perform the extraction. After that, a myriad of studies formulated the extraction task as named entity recognition (NER) (Putthividhya and Hu, 2011; Bing et al., 2012; Ling and Weld, 2021; More, 2016). However, these approaches carry a closed world assumption and therefore can not discover new values of attributes.

With the advent of deep learning, a number of sequence tagging methods were proposed (Kozareva et al., 2016; Huang et al., 2015; Zheng et al., 2018).

These approaches make instead an open world assumption to discover new attribute values. (Huang et al., 2015) applied a BiLSTM-CRF model in a sequence tagging setting. (Zheng et al., 2018) developed an end-to-end tagging model (OpenTag) that benefits from an attention layer (Vaswani et al., 2017) to generate interpretable results. Moreover, (Xu et al., 2019) proposed to encode both attributes and values by using one set of BIO tags to scale up the tagging methods. (Karamanolakis et al., 2020) proposed a taxonomy aware multi-task framework that utilizes the taxonomy of the products to further improve the extraction. (Yan et al., 2021) utilize a hypernetwork (Ha et al., 2017) and Mixture-of-Experts module to parameterize their model with pre-trained attribute embeddings. (Sabeh et al., 2022b) proposed to utilize character level representations to improve the generalization performance of sequence tagging models for extracting brand values from product descriptions. The latest approaches (Wang et al., 2020; Yang et al., 2022; Sabeh et al., 2022a) reformulate the problem as a question answering (QA) task by utilizing BERT (Devlin et al., 2019), which allows them to scale to a large number of attributes. Sequence tagging approaches (Huang et al., 2015; Zheng et al., 2018; Sabeh et al., 2022b) are most relevant to our work because identifying brand names does not require scalability. However, these models are extractive and therefore can not infer brand names which are not directly mentioned in the title or description. They also fail to take product categories into account, which is crucial for brand value identification.

In this work, we adopt a generative approach to identify the brand values from the product descriptions. Our approach allows us to decode brand values that are not directly stated in the text descriptions. Our model is also category-aware, which allows us to effectively take the product categories into account.

3 Proposed Method

As mentioned above, previous methods formalize the brand value identification as a sequence tagging task. These approaches fail to identify brand values that are not explicitly mentioned in the product description. In this work, we tackle the task of brand value identification in a generative manner. More specifically, we propose to fine-tune a generative language model by formulating the brand value identification problem as a sequence-to-sequence generation task.

3.1 Problem Definition

In this section, we formally define the problem of brand value identification from the product description. Given an input product title $t = \{t_1, t_2, \ldots, t_n\}$ where n is the number of tokens in t. We refer to the product category as $c = \{c_1, c_2, \ldots, c_m\} \in C$, where m is the number of tokens in c, and C is a predefined set of categories. The goal of brand value identification is to generate a target sequence $\hat{v}$, which represents the target brand value. For the example in Figure 1a we have:

- t = "Mielle Organics Pomegranate & Honey Moisturizing and Detangling Shampoo, Hydrating Curl Cleanser For Dry, Damaged Type 4 Hair."
- c = "Shampoos"

To generate the value $\hat{v}$ given the product title t and the category c, we formulate the problem as a conditional sequence generation task. Formally, we optimize the model to maximize the conditional log-likelihood $P(v \mid t, c)$ as follows:

$$\hat{v} = \arg\max_{v} P(v \mid t, c)$$

In our implementation, similar to other sequence-to-sequence learning settings (Sutskever et al., 2014), we factorize the log-likelihood into word and sub-word level predictions.

3.2 Language Model Fine-tuning

We employ T5 (Raffel et al., 2020) sequence-to-sequence language model as the backbone of our approach. T5 is a tranformer-based (Vaswani et al., 2017) pre-trained generative language model that maps a given input sequence into an output sequence. The pre-trained T5 model achieves superior performance in many sequence-to-sequence tasks (Qi et al., 2020; Iqbal and Qureshi, 2022). Fine-tuning T5 language model for brand value identification can be done in a similar fashion as for sequence-to-sequence generation tasks, such as machine translation or text summarization, where the model generates a sequence of tokens given the input tokens (Dong et al., 2019; Bao et al., 2020; Xiao et al., 2021).

To make the model aware of the product category c, we propose to concatenate the input title

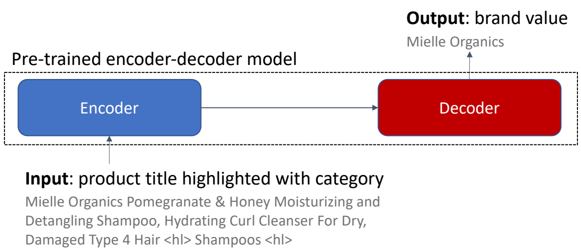


Figure 2: Overview of our generative approach GAVI; it takes category highlighted product title as input and returns the brand value. In this example, the model generates *Mielle Organics* as output.

t and category c into a single input x. After that, we highlight the category in the input. Specifically, following (Chan and Fan, 2019), we introduce a highlight token <hl> to take into account the category c inside the model input x as below:

$$x = \{t_1, t_2, \ldots, t_n, \texttt{<hl>}, c_1, c_2, \ldots, c_m \texttt{<hl>}\}$$

We could also choose not to include and highlight the category in our input. This means that we can also train a generative model that is not category aware by using only the title in our input x:

$$x = \{t_1, t_2, \ldots, t_n\}$$

In our experiments, we investigate and analyse these model variations, but assume the category highlighted title as the default input. We refer to the proposed category aware implementation of the T5 generative model as GAVI in our experiments. Figure 2 shows the overall architecture of our sequence-to-sequence generative approach.

4 Experimental Setup

In this section, we represent the experimental settings of our empirical approach for comparing our generative proposed models with state-of-the-art baselines on the task of brand value identification.

4.1 Datasets

We evaluate our model on a public product dataset[3] for brand value identification (Sabeh et al., 2022b). This dataset comprises over 250k product titles containing more than 50k unique brand values, derived from the Amazon Review Dataset (Ni et al., 2019). Each example consists of product title, product category, and the target brand value for identification.

[3] https://github.com/kassemsabeh/open-brand.

Category	Number of Samples	Average Tokens
Grocery & Gourmet Food	22397	23.22
Toys & Games	63304	21.95
Sports & Outdoors	54214	21.57
Electronics	47870	32.17
Automotive	66837	23.75
Clothing, Shoes & Jewelry	85068	20.75
Pet Supplies	10868	23.72
Cell Phones & Accessories	78564	34.62

Table 1: Detailed statistics of the dataset. We use T5 tokenizer to tokenize the examples.

Category	Train	Val	Test
Grocery & Gourmet Food	15679	2239	4479
Toys & Games	44314	6330	12660
Sports & Outdoors	37951	5421	10842
Electronics	33512	4787	9574
Automotive	45132	6447	12894
Total	176588	25224	50449

Table 2: Statistics of AZ-base dataset with five selected categories.

Table 1 shows the statistical details of the dataset. The dataset contains information about products in eight main categories. The average number of tokens per sample in each category is also shown in Table 1. Following previous work (Sabeh et al., 2022b), we arrange the following setups for benchmark:

- **AZ-base** This split of the dataset contains information about products in five main categories: *Grocery & Gourmet Food, Toys & Games, Sports & Outdoors, Electronics* and *Automotive*. In this dataset, we randomly select 70% of the data for training, 20% for validation, and 10% for testing. The main purpose of this dataset is to evaluate the baseline model performance on the task of brand value identification. The statistics of the AZ-base dataset are provided in Table 2.

- **AZ-zero-shot** In order to evaluate the generalization ability of the models, we divide the AZ-base dataset into another disjoint training and test split with no overlapping brand

Category	Train	Val	Test
Clothing, Shoes & Jewelry	0	0	85068
Pet Supplies	0	0	10868
Cell Phones & Accessories	0	0	78564
Total	0	0	174500

Table 3: Number of samples in AZ-new-cat dataset.

values. The test set of this split contains 8k unique values. None of these values are seen during training. This allows us to evaluate the zero-shot performance of the models.

- **AZ-new-cat** In this benchmark, we test the models ability in identifying brand values from different product categories. In specific, we use the same training set from AZ-base, but we test the model on three new categories of products. None of these categories are present in the training set, as shown in Table 3.

4.2 Implementation Details

All models are implemented using PyTorch[4], and are trained on NVIDIA Tesla V100 GPUs. During training, Adam (Kingma and Ba, 2015) optimizer is applied with initial learning rate $4e^{-5}$. The backbone uses the pre-trained T5-base encoder with 12 layers and 12 heads, which has 220M parameters. The embedding dimension is 768, while the maximal input length is set to 512. The batch size is set to 32. All hyper-parameters are chosen optimally based on the performance on the validation set of our dataset. We fine-tune the model on the training set for 10 epochs, and perform early stopping if there is no improvement in the loss on the validation set for 3 epochs. We report our final results on the test where we perform beam search of size four.

4.3 Evaluation Metrics

Following the literature (Xu et al., 2019; Yan et al., 2021; Wang et al., 2020), we use Precision (P), Recall (R), and F_1 as evaluation metrics. We compute these metrics based on the number of true positives (TP), false positives (FP), and false negatives (FN) of our predictions. We use Exact Match (Rajpurkar et al., 2016) criteria in our evaluations, where the full predicted sequence should match the ground truth.

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F_1 = 2 \times \frac{P \times R}{P + R}$$

4.4 Compared Models

We compare the following models on the task of brand value identification:

BiLSTM-CRF (Huang et al., 2015) applies a BiLSTM followed by a CRF layer to model the dependency of the predicted tags.

[4]https://pytorch.org/

OpenTag (Zheng et al., 2018) introduces a self-attention layer between the BiLSTM and CRF to highlight important features in the input. OpenTag is considered as the pioneer sequence tagging model for attribute value extraction.

OpenBrand[5] (Sabeh et al., 2022b) leverages a CNN encoder to generate character level representations and improve the generalization performance. OpenBrand achieves state-of-the-art results on the brand value extraction task.

T5 The base generative model of ours that is fine-tuned on the training dataset. T5 is not category-aware as it only uses the title in the input.

GAVI Our proposed category-aware generative model. Our model uses category-highlighted inputs to identify the brand values, as described in Section 3.2.

5 Experimental Results

In this section, we conduct a series of experiments under various settings to evaluate our proposed approach.

5.1 Baseline Comparison Results

In this experiment, we compare the performance of our proposed models with the baseline models, as mentioned in Section 4.4, on the AZ-base dataset. Table 4 reports the evaluation results of the compared models on the five categories of AZ-base. We can observe that GAVI consistently outperforms other baselines across all categories of products. The overall improvement in F_1 score is up to 4.1% compared to OpenBrand. One interesting observation is that both T5 and GAVI outperform the other baselines in terms of F_1 score. The main reason is that both models are generative, meaning that they can identify brand values that are not mentioned in the title. On the other hand, sequence tagging approaches fail to extract such values.

We also notice that, in general, GAVI outperforms the base T5 model in all categories of products. This is mainly because our model is category-aware and is able to learn category-specific embeddings that are more suitable for the identification task. Another key observation is that the performance of GAVI depends on product categories. For

[5]We only compare our model with the CNN version of OpenBrand as it was shown to have better performance by the authors.

Category	Models	P	R	F_1
Grocery & Gourmet Food	BiLSTM-CRF	74.9	66.0	70.2
	OpenTag	76.0	65.4	70.3
	OpenBrand	77.5	75.4	76.4
	T5	76.5	75.9	76.2
	GAVI	**79.3**	**76.4**	**77.8**
Toys & Games	BiLSTM-CRF	78.9	70.5	74.5
	OpenTag	79.1	70.3	74.5
	OpenBrand	**81.3**	72.0	76.4
	T5	79.7	76.6	78.1
	GAVI	80.3	**77.2**	**78.7**
Sports & Outdoors	BiLSTM-CRF	84.1	75.4	79.5
	OpenTag	84.9	75.0	79.6
	OpenBrand	86.1	77.3	81.5
	T5	82.1	81.5	81.8
	GAVI	**88.1**	**82.3**	**85.1**
Electronics	BiLSTM-CRF	87.8	81.5	84.5
	OpenTag	89.2	79.6	84.2
	OpenBrand	89.7	80.5	84.9
	T5	87.9	81.5	87.8
	GAVI	**90.1**	**88.5**	**89.3**
Automotive	BiLSTM-CRF	90.9	85.0	87.9
	OpenTag	91.6	84.6	87.9
	OpenBrand	**91.8**	85.4	88.5
	T5	90.4	90.5	90.4
	GAVI	91.4	**91.3**	**91.3**

Table 4: Performance comparison between different models on the AZ-base dataset.

Model	P	R	F_1
OpenTag	53.80	33.82	41.53
OpenBrand	55.61	35.46	43.44
T5	67.28	47.90	55.95
GAVI	**70.10**	**53.31**	**60.55**

Table 5: Results on zero-shot brand values.

example, the gain in recall R in the *Electronics* category (7%) is much higher than the gain in the *Grocery & Gourmet Food* category (1%). By analyzing the errors in the *Grocery & Gourmet Food* category, we discovered that there are certain amount of false negatives in the test set, where the outputs of the model are actually correct, but the labels are wrong. For example, given the following title: "Organo Gold Organic Green Tea (4 Boxes)", the model correctly extracts "Organo Gold" as the brand, but the ground truth is "Organic Green Tea".

5.2 Results of Discovering New Brand Values

We conduct zero-shot extraction experiments to evaluate the generalization performance of our model on unseen brand values. The results on the zero-shot dataset are reported in Table 5. We exclude the BiLSTM-CRF model from this experiment as its not capable to generalize well to new values. It can be seen that our generative models achieve much better results than OpenBrand and OpenTag. For example, the F_1 metric of GAVI significantly increases by 17.1% compared with OpenBrand over all categories in the dataset. This is because generative models use the T5 transformer-based (Vaswani et al., 2017) architecture, which have been shown to outperform the BiLSTM-CRF architecture in zero-shot settings (Wang et al., 2020).

From Table 5, we can also observe that GAVI outperforms the base T5 model on the zero-shot extractions (e.g., by 4.6% F_1 score). This is because knowledge about the category allows the model to exploit similarities across product categories resulting in a better overall performance. However, it is evident that the overall performance of the models is worse as compared to the main results in Table 4. This is expected, as there are no examples from the zero-shot brand values in our training set.

Category	Models	P	R	F_1
Clothing, Shoes, & Jewelry	BiLSTM-CRF	58.5	42.2	49.0
	OpenTag	60.3	43.5	50.5
	OpenBrand	64.5	45.2	53.2
	T5	64.2	**55.9**	57.4
	GAVI	**64.5**	55.8	**59.8**
Pet Supplies	BiLSTM-CRF	55.0	37.3	44.5
	OpenTag	53.9	38.9	45.2
	OpenBrand	58.2	38.5	46.3
	T5	**64.8**	**51.5**	**57.4**
	GAVI	63.6	49.3	55.6
Cell Phones & Accessories	BiLSTM-CRF	80.1	68.0	73.5
	OpenTag	78.3	67.4	72.4
	OpenBrand	85.2	67.8	75.5
	T5	85.4	81.5	83.4
	GAVI	**85.5**	**81.7**	**83.6**

Table 6: Performance comparison between models on the AZ-new-cat dataset.

5.3 Results on New Categories

To examine the models ability in generalizing to brand values in new categories, we conduct a set of experiments using the AZ-new-cat benchmark. In these experiments, we train the models on the training set of the AZ-base dataset and evaluate them on three new product categories: *Clothing, shoes, & Jewelry*, *Pet Supplies*, and *Cell Phones & Accessories*. We report the results of our experiments in Table 6. It can be seen that GAVI outperforms all

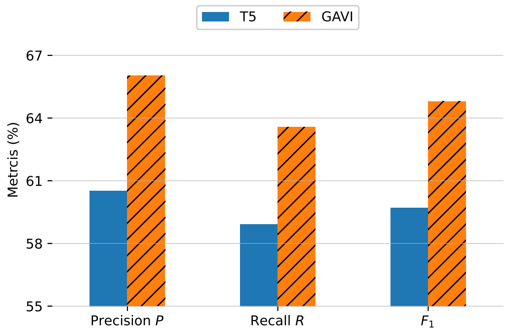


Figure 3: Performance comparison of T5 and GAVI on instances of AZ-base where the brands are not explicitly mentioned in the product description.

the compared baselines. The increase in F_1 score is up to 9.3% as compared to OpenBrand. These results demonstrate the models ability in generalizing to new domains in real-world scenarios.

There are several interesting observations in Table 6. First, the performance of T5 and GAVI is close. This is because using new categories that are not seen during training does not benefit our category-aware implementation. The model is not able to generate category-specific token embeddings at inference time, as they are new categories that are unseen during training. Second, and inline with previous works (Sabeh et al., 2022b), the results on the *Cell Phones & Accessories* category are significantly better than other categories for all compared models. This is because many of the brands in the *Cell Phones & Accessories* category are also present in the *Electronics* category of the training set (e.g., "LG").

5.4 Results on Implicit Brands Examples

We further conduct a set of experiments on AZ-base to analyze the performance of the models on the instances where the brand value is not explicitly mentioned in the product description. We refer to those examples as *implicit examples* (e.g., the product in Figure 1b). First, we separate the implicit examples in the test set of AZ-base. This resulted in 9k implicit examples. Then, we fine-tune the models on the training set of AZ-base and test them on these implicit examples. Figure 3 shows the evaluation results of T5 and GAVI on the implicit examples. Note that we do not include the other extractive baselines in this experiment as they are not able to extract those implicit brands.

GAVI achieves 64.9% F_1 score on the implicit examples. This indicates the effectiveness of our approach compared to sequence tagging baselines, which are incapable of performing the extraction. In addition, GAVI significantly outperforms the base T5 model in all compared metrics. This clearly indicates that taking the categories into consideration during the generation results in better overall performance.

5.5 Examples of Extracted Brand Values

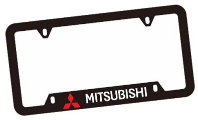
Figure 4 shows examples of product titles and brand values extracted by OpenBrand or GAVI. GAVI is able to identify brands that are not explicitly mentioned in the title: in Figure 4a, "Frame pro" is the valid brand for this product. OpenBrand, which is an extractive sequence tagging model, fails to detect this value. Instead, it extracts "Mitsubishi" as the brand. While GAVI successfully generates "Frame pro" as the correct brand value. In Figure 4b, OpenBrand erroneously extracts "Fun" as the brand for a *Toys & Games* product; on the other hand, GAVI, which considers the product category and textual context, generates the correct brand for this product. Also, in the example of Figure 4c, the model was able to correctly extract the brand value, even though it was mentioned incorrectly in the title.

6 Conclusions and Future Work

Brand value identification is a crucial task in many real-world e-commerce applications. In this work, we propose a novel generative approach for brand value identification. In particular, we employ T5 language model as the backbone of our sequence-to-sequence approach. We infuse category information into the model by highlighting the product categories inside the input. In contrary to previous extractive approaches, our generative method allows us to identify brands beyond the strings mentioned in the product description. Experimental evaluations on public datasets demonstrate the effectiveness of the proposed approach.

We plan to investigate other sequence-to-sequence language models such as BART (Lewis et al., 2020) and GPT (Brown et al., 2020). Also, improving the brand coverage and dealing with the false negatives in the dataset is one of the future directions.

Title: Mitsubishi 3000GT License Plate Frame (Zince Metal)



Brand: Frame pro
OpenBrand = "Mitsubishi"
GAVI = "Frame pro"

(a)

Title: Fun Fire Truck Pinata Personalized

Brand: Personalized Pinatas
OpenBrand = "Fun"
GAVI = "Personalized Pinatas"

(b)

Title: Fisher-Price Thomas&quotFriends Take-n-play

Brand: Thomas & Friends
OpenBrand = "Fisher-Price"
GAVI = "Thomas & Friends"

(c)

Title: White Chocolate Caramel Gourmet Popcorn Kelly

Brand: Kelly
OpenBrand = "Kelly"
GAVI = "Kelly"

(d)

Figure 4: Examples of extracted brand values from OpenBrand and GAVI.

Limitations

In this paper, we introduce a novel generative approach for brand value identification from product descriptions. The input to our models is limited to up to around 500 tokens, and the same approach can not be easily applied to longer product descriptions. As far as languages are concerned, the models developed here are English only. To adapt our work to other languages, we need e-commerce datasets to train and evaluate the models in those languages. Also, our models assume that the brand values can always be identified from the context of the product descriptions. We do not consider the case where the context does not include any applicable brand value (i.e., negative values). As future work, we will extend the model and datasets to deal with those negative samples.

References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Lidong Bing, Tak-Lam Wong, and Wai Lam. 2012. Unsupervised extraction of popular product attributes from web sites. In *Information Retrieval Technology*, pages 437–446, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

Henrieta Hrablik Chovanová, Aleksander Ivanovich Korshunov, and Dagmar Babčanová. 2015. Impact of brand on consumer behavior. *Procedia Economics and Finance*, 34:615–621. International Scientific Conference: Business Economics and Management (BEM2015).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rayid Ghani, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. 2006. Text mining for product attribute extraction. *SIGKDD Explor. Newsl.*, 8(1):41–48.

Vishrawas Gopalakrishnan, Suresh Parthasarathy Iyengar, Amit Madaan, Rajeev Rastogi, and Srinivasan

Sengamedu. 2012. Matching product titles using web-based enrichment. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 605–614, New York, NY, USA. Association for Computing Machinery.

David Ha, Andrew M. Dai, and Quoc V. Le. 2017. Hypernetworks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Touseef Iqbal and Shaima Qureshi. 2022. The survey: Text generation models in deep learning. *Journal of King Saud University - Computer and Information Sciences*, 34(6, Part A):2515–2528.

Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. 2020. TXtract: Taxonomy-aware knowledge extraction for thousands of product categories. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8489–8502, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Zornitsa Kozareva, Qi Li, Ke Zhai, and Weiwei Guo. 2016. Recognizing salient entities in shopping queries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 107–111, Berlin, Germany. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiao Ling and Daniel Weld. 2021. Fine-grained entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):94–100.

Ajinkya More. 2016. Attribute extraction from product titles in ecommerce. *CoRR*, abs/1608.04670.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Duangmanee Putthividhya and Junling Hu. 2011. Bootstrapped named entity recognition for product attribute extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Kassem Sabeh, Mouna Kacimi, and Johann Gamper. 2022a. Cave: Correcting attribute values in e-commerce profiles. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 4965–4969, New York, NY, USA. Association for Computing Machinery.

Kassem Sabeh, Mouna Kacimi, and Johann Gamper. 2022b. OpenBrand: Open brand value extraction from product descriptions. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 161–170, Dublin, Ireland. Association for Computational Linguistics.

Umer Shahzad, Salman Ahmad, Kashif Iqbal, Muhammad Nawaz, and Saqib Usman. 2014. Influence of brand name on consumer choice & decision. *IOSR Journal of Business and Management*, 16:72–76.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. Generative language models for paragraph-level question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 670–688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Damir Vandic, Jan-Willem van Dam, and Flavius Frasincar. 2012. Faceted product search powered by the semantic web. *Decision Support Systems*, 53(3):425–437.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D. Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. 2020. Learning to extract attribute value from product via question answering: A multi-task approach. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 47–55. ACM.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. 2019. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223, Florence, Italy. Association for Computational Linguistics.

Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. 2021. AdaTag: Multi-attribute value extraction from product profiles with adaptive decoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4694–4705, Online. Association for Computational Linguistics.

Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2022. MAVE: A product dataset for multi-source attribute value extraction. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 1256–1265. ACM.

Yi Zhang et al. 2015. The impact of brand image on consumer behavior: A literature review. *Open journal of business and management*, 3(01):58.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1049–1058. ACM.