Knowledge-Enhanced Multi-Label Few-Shot Product Attribute-Value Extraction

Jiaying Gong
Virginia Polytechnic Institute and State University
Blacksburg, U.S.
gjiaying@vt.edu

Wei-Te Chen
Rakuten Institute of Technology, Rakuten Group Inc.
Boston, U.S.
weite.chen@rakuten.com

Hoda Eldardiry
Virginia Polytechnic Institute and State University
Blacksburg, U.S.
hdardiry@vt.edu

ABSTRACT

Existing attribute-value extraction (AVE) models require large quantities of labeled data for training. However, new products with new attribute-value pairs enter the market every day in real-world e-Commerce. Thus, we formulate AVE in multi-label few-shot learning (FSL), aiming to extract unseen attribute value pairs based on a small number of training examples. We propose a Knowledge-Enhanced Attentive Framework (KEAF) based on prototypical networks, leveraging the generated label description and category information to learn more discriminative prototypes. Besides, KEAF integrates with hybrid attention to reduce noise and capture more informative semantics for each class by calculating the label-relevant and query-related weights. To achieve multi-label inference, KEAF further learns a dynamic threshold by integrating the semantic information from both the support set and the query set. Extensive experiments with ablation studies conducted on two datasets demonstrate that our proposed model significantly outperforms other SOTA models for information extraction in few-shot learning.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

attribute value extraction; multi-label few-shot learning

ACM Reference Format:
Jiaying Gong, Wei-Te Chen, and Hoda Eldardiry. 2023. Knowledge-Enhanced Multi-Label Few-Shot Product Attribute-Value Extraction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3583780.3615142

1 INTRODUCTION

Product attribute value pairs are important for e-Commerce because platforms make product recommendations for customers based on the key attribute-value pairs and customers use attributes to compare products and make purchases. Existing studies on AVE based on neural networks view AVE as sequence labeling [13, 34], question-answering [24, 28] or multi-modal fusion problems [15, 41]. These supervised-learning models are well trained to classify attribute-value pairs when large quantities of labeled data are available for training. Even the most current open mining model needs a few attribute-value seeds and iterative training for weak supervision [38]. However, new products with new attribute-value pairs enter the market every day in real-world e-Commerce platforms. It is difficult, time-consuming and costly to manually label large quantities of new products profiles for training. Besides, with the appearance of new attribute-value pairs, the class distribution becomes long-tailed, where a subset of the labels have many samples, while majority of the labels have only a few samples.

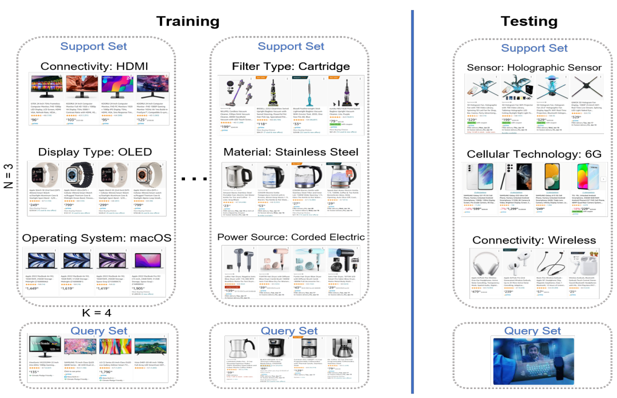

Figure 1: An example of multi-label few-shot product attribute-value extraction task.

We formalize AVE as a multi-label FSL problem, aiming to extract structured product information from unstructured profiles with limited training data. We take the common head labels data for training and the limited tail labels data for testing, and there is no overlap of classes between training set and testing set shown in Figure 1. Recent methods on multi-label FSL have made great progress in CV [1, 25] and NLP [10, 17, 39]. Among these methods, prototypical network [26] has been proved to be powerful and potential. However, different from AVE in e-Commerce, these models (1) explore only label tags as auxiliary information, (2) still have noise when learning prototypes, and (3) require further data or additional models to learn the threshold for label numbers prediction.

To address the above challenges, we propose a Knowledge - Enhanced Attentive Framework (KEAF) for product AVE. The main contributions of KEAF consist of three parts. (1) To the best of our knowledge, we are the first to formulate AVE as a multi-label FSL task to tackle the problem of limited training data for long-tailed datasets. Unlike open mining models, KEAF does not require the attribute-value pairs exactly appear in the product profiles. (2)

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615142

By leveraging both the label description generated by a generator and the category information as the auxiliary information to obtain more discriminative prototypes, KEAF can not only avoid the issue that different attribute-value pairs share the identical prototype for 1-shot learning, but also alleviate ambiguity by obtaining both label and category relevant information. The hybrid attention mechanism also helps reduce the noise and capture more informative semantics from the support set by calculating both the label-relevant and query-related weights. (3) To achieve multi-label inference, a dynamic threshold is learned during the training stage by integrating the semantic information from support and query sets. The adaptive threshold does not require additional training data or based on additional models. Extensive experimental results on two datasets show that our proposed model KEAF significantly outperforms other existing information extraction models for AVE.

2 RELATED WORKS

Early works on AVE use a domain-specific dictionary and rule-based methods to identify attribute value pairs [8, 19, 23, 32]. Then, sequence labeling models [13, 21, 34, 40], question answering-based models [24, 28, 33], multi-modal models [15, 29, 41], extreme multi-label models [3, 4] and open mining models [38] are trained for AVE. However, these approaches require large quantities of labeled data for training or iterative training for weak supervision. Most works of FSL focus on single-label classification [2, 6, 9, 14, 18]. However, one product may have multiple attribute value pairs for AVE task. Early works on multi-label FSL depends on a known structure of the label spaces [22] and label set operations [1]Then, prototypical networks [27] are revised for multi-label cases by learning a shared embedding space [37], grouping samples multiple times [25], and learning local features with different labels [35]. Attention mechanisms [11] and label information [10, 17, 39] are considered to differentiate prototypes. Different from these approaches, we leverage both label and category for product AVE in e-Commerce.

3 METHODOLOGY

3.1 Problem Definition

Given a set of training classes Y_{train} and testing classes Y_{test}, where $Y_{train} \cap Y_{test} = \emptyset$. The model is trained with numerous samples from Y_{train}, and it can quickly adapt to Y_{test} with few labeled data. Each training episode involves a support set $S = \left\{(x_i, y_i)\right\}_{i=1}^{N_s}$ and a query set $Q = \left\{(x_i, y_i)\right\}_{i=1}^{N_q}$, where S usually includes K samples (K-shot) for each of N labels (N-way). In contrast to the single label N-way-K-shot setting [30], multi-label FSL allows that each single sample can have multiple labels simultaneously. There are N total classes, and each class has at least K samples (with at least one label appearing less than K times if any samples are removed) because we can not guarantee each label appears exactly K times while each sample has multiple labels. The input data for each product x is a tuple $< t, d, l, c >$, where t is the product title, d is the product description, l is the label description, and c is the product category. The input label is a vector $y = \left\{y_1, y_2, \ldots, y_N\right\}$, where $y \in \left\{0, 1\right\}$, indicating whether the product has the label or not, and N is the total number of classes. The outputs are attribute-value pairs.

3.2 Multi-label Few-Shot Data Sampling

Multi-label few-shot data sampling includes data splitting, data balancing and data sampling. We first reconstruct the dataset by splitting data based on upper thresholds t_u and lower thresholds t_l, learned from the frequency of class labels to guarantee that $Y_{train} \cap Y_{test} = \emptyset$. We filter the dataset by discarding the samples with the label count below t_l or above t_u, updating the label dictionary and its corresponding samples.To guarantee that the shot $K_S + K_Q >= 10$ for FSL, the filtering process is done iterately until class number N is fixed. Then, we balance the dataset by randomly dropping single-label data to achieve a similar size with multiple-label data. To approximately conduct N-way-K-shot learning, we follow [10] to construct query and support sets for each episode. Details of multi-label few-shot data sampling are shown in Algorithm 1.

Algorithm 1: Multi-label Few-shot Data Sampling

```
Input  : Dataset X, label set Y, support shot K_S, query shot K_Q,
         upper threshold t_u and lower threshold t_l
Output : Support set S, query set Q, query label set Q_L
Initialize S = {}, Q = {}, Q_L = [ ] and Dict {label : count}
while len(Y) is not fixed do
    if Count(Y_{X_{i,j}}) > t_u or Count(Y_{X_{i,j}}) <= t_l then
        X.remove(X_{i,j})
    Y.update(X)
    if get_class(X_{i,j}) ∉ Y then
        X.remove(X_{i,j})
data_balancing(X)
for i in Enumerate(Y) do
    indices = Random(Y_i, K_S + K_Q); count = 0
    for j in indices do
        if count < K_Q then
            Q.update(X_{i,j}); Q_L.append(Y_{i,j})
        else
            if any Dict[Y_{i,j}] < K_S then
                S.update(X_{i,j}); Update Dict
        count++
return S, Q, Q_L
```

3.3 Knowledge-Enhanced Attentive Framework

In this section, we introduce the overview of KEAF in Figure 2.

3.3.1 Contextual Representations. Labels for AVE tasks are attribute value pairs such as 'wallet type: long wallet', which may lose contextual information due to the simple format. To achieve more information related to labels, we adopt GPT-2 [20]as the text generator to generate a detailed description for the attribute-value pairs. We adopt a pre-trained language model BERT [5]as the product input encoder to generate the contextual representation. We construct a string [CLS;c;SEP;t;SEP;d] by concatenating product category, title and description as the input.The output representation for the product input r_i and label input l_i is:

$$r_i = tanh(W \cdot f_{\emptyset}(c_i, t_i, d_i) + b), \quad l_i = tanh(W \cdot f_{\emptyset}(g_{\emptyset}(l_i)) + b) \quad (1)$$

where $f_{\emptyset}$ is BERT encoder, $g_{\emptyset}$ is GPT-2 generator, c is category, t is title, d is description and l is 'attribute is value' label information.

3.3.2 Label-Enhanced Prototypical Network. In Figure 2, we adopt prototypical networks [26] to get the original prototype of each attribute-value pair by averaging the embedding of support samples.

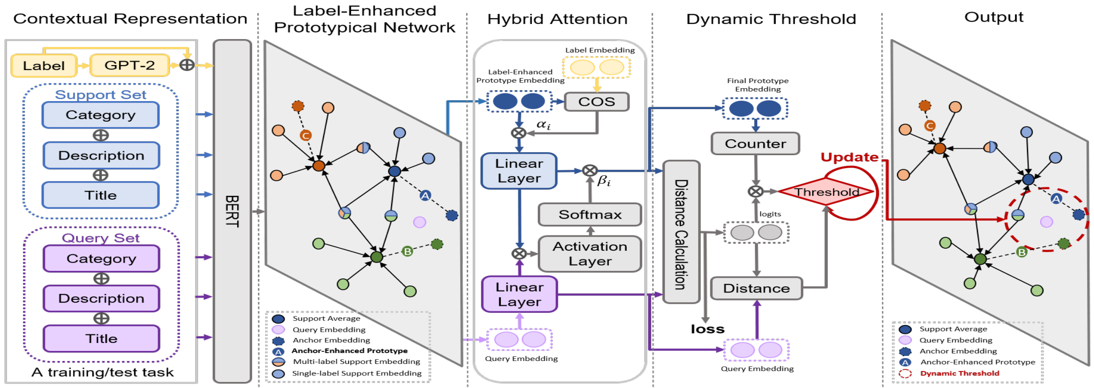

Figure 2: The overview of our proposed KEAF framework.

However, different labels may share the same support samples in multi-label settings, resulting in severe ambiguity. To emphasize the difference between prototypes and reduce such ambiguity, we leverage label descriptions generated by GPT-2 [20] to fully express the semantic information for attribute-value pairs and help learn more representative prototypes. Label has shown significant effect on learning more discriminative prototypes [10, 17, 39]. Thus, we combine the label with the average of support samples to compute a label-enhanced prototype c_i with an interpolation factor η:

$$c_i = \eta \times E(y_i) + (1 - \eta) \times \frac{1}{K_i} \sum_{j=1}^{K_i} E(x_i^j) \tag{2}$$

where $E(\cdot)$ is the BERT encoder, y_i is the label description, $x_i^j \in \{x | (x, Y) \in S \wedge y_i \in Y\}$ is the support sample labeled with y_i, and K_i is shot number. The combination of label description and support embedding helps the prototypes better separated from each other.

3.3.3 Hybrid Attention. The aim of hybrid attention is to select more informative instances by retaining attribute-value relevant information while eliminating the negative effect triggered by the noise. As shown in the third stage in Figure 2, we first capture the similarity weight α_i in the label by calculating the semantic similarity between the label-enhanced prototype embedding c_i from Equ 2 and the attribute-value description embedding l_i from Equ 1:

$$\alpha_i = cos(c_i, l_i), \quad \hat{c}_i = \alpha_i \times c_i \tag{3}$$

where $cos(\cdot)$ is the cosine similarity and $\hat{c}_i$ gets the class-relevant information. To further capture informative semantics from query-related instances and reduce the noise, we apply the instance-level attention, where each instance has a different importance factor β_i:

$$\beta_i = \frac{exp(L(\hat{c}_i) \times L(E(x_i^q)))}{\sum_{i'=1}^{K} exp(L(\hat{c_{i'}}) \times L(E(x_{i'}^q)))}, \quad \hat{r}_i = \beta_i \times \hat{c}_i \tag{4}$$

where $L(\cdot)$ is the linear layer, $E(\cdot)$ is encoder from Equ. 1, x_i^q represents the query instance and $\hat{r}_i$ is the final prototype. Now, $\hat{r}_i$ contains label-relevant semantic information and it can be closer to the instances with features more related to queries.

3.3.4 Dynamic Threshold. In Figure 2, we train the threshold τ by integrating the semantic information from both the support and query sets. The thresholding function $T(\cdot)$ is calculated by the production of query label counter $\varphi(x_i^q)$ with the relevance score between the final prototype $\hat{r}_i$ in Equ. 4 and query instance r_i^q generated from Equ. 1. The number of query labels is estimated by averaging the number of support labels of support instance x_i:

$$\tau = T(\varphi(x_i^q), S) = \frac{1}{N \times K} \sum_{x_i \in X} \varphi(x_i) \odot d(\hat{r}_i, r_i^q) \tag{5}$$

where S is the support set, N and K denotes N-way-K-shot, $\varphi(\cdot)$ represents the label counter, $\odot$ is element-wise production, and $d(\cdot)$ is the distance function. The threshold is dynamically updated for each training epoch. In testing, the framework predicts the query label set Y_i^q by comparing the distance d_i^q with the threshold τ:

$$Y_i^q = \{y_i^q | d_i^q < \tau, y_i^q \in Y\} \tag{6}$$

The final threshold for the testing phase is chosen by the threshold value that has the best performance in the evaluation phase. The model is trained by repeatedly sampling training episodes from Y_{train} with support set S and query set Q. The model parameters are updated using the following binary cross entropy (BCE) loss:

$$\mathcal{L} = \sum_{I \in Q} \sum_{i=1}^{N} y_i^I \cdot log\sigma(q_i^I) + (1 - y_i^I) \cdot log(1 - \sigma(q_i^I)) \tag{7}$$

where Q is query shot, N is N-way, $\sigma(\cdot)$ is the sigmoid function and y_i^I represents the ground truth.

Table 1: Comparison with other multi-label FSL datasets.

Dataset	Train/Val/Test #Instance	Train/Val/Test #Label	Multi-label Percentage
MS-COCO [16]	97,600/-/24,400	64/-/16	38.81%
FewAsp [12]	40,960/10,240/12,800	64/16/20	63.5%
TourSG [31]	19,351/1,600/4,800	68/17/17	18.13%
StanfordLU [7]	3,517/2,512/2,009	14/10/8	16.57%
MAVE [36]	**29,458/-/2,049**	**45/-/17**	**45.25%**
Ours	**477,166/-/6,421**	**23/-/14**	**43.38%**

4 EXPERIMENTS

4.1 Experimental Setup

We evaluate our model over two datasets: a large e-Commerce platform in Japanand MAVE [36]. The dataset statistics is shown in

Table 2: Experimental results (%) of multi-label few-shot learning on an in-house E-Commerce dataset.

	1-shot						5-shot					
Model	Mac-P	Mac-R	Mac-F1	Mic-P	Mic-R	Mic-F1	Mac-P	Mac-R	Mac-F1	Mic-P	Mic-R	Mic-F1
Siamese [14]	12.52	30.84	16.07	12.11	30.23	17.18	22.16	25.76	21.75	21.38	24.55	22.76
MTB [2]	13.00	36.86	17.70	12.77	36.76	18.87	10.16	**98.89**	18.06	10.12	**98.51**	18.35
Proto_BERT [26]	24.71	39.73	28.00	**30.44**	41.82	34.78	30.71	40.39	32.81	**32.85**	42.55	36.86
HCRP [9]	21.96	39.03	23.36	23.33	35.39	27.65	18.90	86.08	28.71	16.25	84.68	27.21
FAEA [6]	22.28	73.01	31.30	19.74	72.66	30.92	23.73	77.05	33.78	22.23	77.99	34.47
SimpleFS [18]	14.75	69.33	23.12	18.08	72.80	28.93	16.81	62.65	25.16	21.59	66.69	32.55
KEAF w/o att	24.79	**73.57**	34.43	24.54	**75.35**	36.92	26.39	73.50	36.69	26.01	75.08	38.48
KEAF	**26.59**	69.10	**35.55**	26.38	69.00	**37.88**	**34.54**	66.96	**42.97**	32.47	63.63	**42.91**

Table 3: Ablation result over components in 1-shot learning setting on in-house E-Commerce and MAVE datasets.

	In-house E-Commerce						MAVE					
Model	Mac-P	Mac-R	Mac-F1	Mic-P	Mic-R	Mic-F1	Mac-P	Mac-R	Mac-F1	Mic-P	Mic-R	Mic-F1
w/o anchor weight	20.93	37.16	24.8	28.63	44.71	34.57	26.45	35.48	25.64	33.22	32.00	32.05
w/o generator	24.59	47.75	29.32	27.59	49.27	35.25	28.56	32.70	26.77	40.29	33.98	36.08
w/o threshold	24.64	44.66	29.37	**28.68**	48.26	35.57	**32.99**	**61.88**	**38.22**	33.45	**66.77**	44.27
w/o category	23.76	56.62	30.98	27.34	60.32	37.38	18.82	23.83	17.88	35.19	21.66	26.27
w/o attention	24.79	**73.57**	34.43	24.54	**75.35**	36.92	32.42	47.89	33.52	34.58	48.39	39.71
KEAF (All)	**26.59**	69.10	**35.55**	26.38	69.00	**37.88**	31.38	50.92	34.47	37.65	55.54	**44.46**

Table 4: Results of F1 score (%) on MAVE dataset.

Model	1-shot		5-shot	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Siamese [14]	15.83	18.55	28.29	28.15
MTB [2]	17.36	18.26	20.62	20.76
Proto_BERT [26]	30.09	36.79	33.56	39.46
HCRP [9]	18.55	19.31	16.11	16.58
FAEA [6]	16.38	16.87	16.63	17.11
SimpleFS [18]	26.63	30.43	17.44	23.05
KEAF w/o att	33.52	39.71	**38.22**	44.27
KEAF	**34.47**	**44.46**	36.40	**44.33**

Table 1. We compare KEAF with SOTA few-shot IE approaches, models without label information: Siamese [14], Proto [26], MTB [2], and models with label information: HCPR [9], FAEA [6] and SimpleFSRE [18]. For evaluation, we use both Micro(Mic-) and Macro(Mac-) Precision, Recall and F1. The max length of input is 512 for project and 32 for anchor. The dimension size is 768. We vary the label interpolation factor η in $\{0.1, 0.5, 1.0\}$ and the optimal anchor weight is selected. Our model is implemented on PyTorch and optimized with AdamW optimizer. The learning rate is 10^{-5} with weight decay 10^{-6}. The batch size is 1 and dropout rate is 0.2. The experiments are conducted on Nvidia A100 GPU with 80G GPU memory.

4.2 Results and Discussions

4.2.1 Main Results. The results of multi-label FSL are shown in Table 2 and Table 4. We observe: (1) KEAF significantly outperforms other baselines on both macro and micro F1 in 1-shot and 5-shot settings. These results reveal that KEAF better learns the prototypes and captures the informative semantics. (2) On the in-house E-Commerce dataset, models using label semantics improve the performance more in 1-shot than 5-shot setting. This is consistent with our expectations that adding label information helps reduce ambiguity. On MAVE, baselines using labels even have a worse performance. We conjecture that the original labels in MAVE are too simple to learn anything, and they even cause noise. In KEAF, we generate a more detailed label description for better integrating the label and reducing the noise, resulting in the best performance among all models. (3) MTB and other baselines demonstrate good results only on Recall and bad results on Precision. For AVE, low precision means lots of human efforts are needed to manually remove the extracted non-relevant attribute-value pairs. A possible reason is that a very large threshold is learned on these baselines, trying to predict as many labels as possible and resulting in a very large recall value. In contrast, KEAF better learns the threshold and balance the Precision and Recall, resulting in the highest F1 score.

4.2.2 Ablation Study. To verify the effectiveness of each component in KEAF, we conduct the 1-shot ablation study in Table 3. We observe: (1) Fusing anchor to the prototypes results in a large performance improvement because the label helps discriminate prototypes and reduce ambiguity. (2) Generating a more detailed label description helps improve the performance more on MAVE than on the in-house E-Commerce dataset. We conjecture that MAVE is an English dataset and English GPT-2 is well trained than Japanese GPT-2, resulting in a more accurate label description on MAVE. (3) Category information shows vital importance on MAVE. We think that AV pairs are from different categories and adding the category can better separate the prototypes. (4) Using the attention can improve the performance by reducing the noise to some extent.

5 CONCLUSION

In this paper, we formulate attribute value extraction task in few-shot learning to solve the long-tailed data problem and limited training data for new products. We propose a Knowledge-Enhanced Attentive Framework for product AVE. We design a label-enhanced prototypical network with the hybrid attention to alleviate ambiguity and noise, and capture more informative semantics. We train a dynamic threshold to achieve multi-label inference. Results demonstrate that KEAF outperforms other SOTA IE models significantly.

REFERENCES

[1] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. 2019. Laso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6548–6557.

[2] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2895–2905. https://doi.org/10.18653/v1/P19-1279

[3] Wei-Te Chen, Yandi Xia, and Keiji Shinzato. 2022. Extreme Multi-Label Classification with Label Masking for Product Attribute Value Extraction. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. 134–140.

[4] Zhongfen Deng, Wei-Te Chen, Lei Chen, and Philip S. Yu. 2022. AE-smnsMLC: Multi-Label Classification with Semantic Matching and Negative Label Sampling for Product Attribute Value Extraction. In *2022 IEEE International Conference on Big Data (Big Data)*. 1816–1821. https://doi.org/10.1109/BigData55660.2022.10020304

[5] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Chunliu Dou, Shaojuan Wu, Xiaowang Zhang, Zhiyong Feng, and Kewen Wang. 2022. Function-words Adaptively Enhanced Attention Networks for Few-Shot Inverse Relation Classification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 2937–2943. https://doi.org/10.24963/ijcai.2022/407 Main Track.

[7] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Saarbrücken, Germany, 37–49. https://doi.org/10.18653/v1/W17-5506

[8] Rayid Ghani, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. 2006. Text Mining for Product Attribute Extraction. *SIGKDD Explor. Newsl.* 8, 1 (jun 2006), 41–48. https://doi.org/10.1145/1147234.1147241

[9] Jiale Han, Bo Cheng, and Wei Lu. 2021. Exploring Task Difficulty for Few-Shot Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2605–2616. https://doi.org/10.18653/v1/2021.emnlp-main.204

[10] Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. 2021. Few-shot learning for multi-label intent detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13036–13044.

[11] Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. 2021. Multi-Label Few-Shot Learning for Aspect Category Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 6330–6340. https://doi.org/10.18653/v1/2021.acl-long.495

[12] Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. 2021. Multi-Label Few-Shot Learning for Aspect Category Detection. 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)*. Association for Computational Linguistics, Online, 6330–6340. https://doi.org/10.18653/v1/2021.acl-long.495

[13] Mayank Jain, Sourangshu Bhattacharya, Harshit Jain, Karimulla Shaik, and Muthusamy Chelliah. 2021. Learning Cross-Task Attribute - Attribute Similarity for Multi-task Attribute-Value Extraction. In *Proceedings of the 4th Workshop on e-Commerce and NLP*. Association for Computational Linguistics, Online, 79–87. https://doi.org/10.18653/v1/2021.ecnlp-1.10

[14] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2. Lille, 0.

[15] Rongmei Lin, Xiang He, Jie Feng, Nasser Zalmout, Yan Liang, Li Xiong, and Xin Luna Dong. 2021. PAM: understanding product images in cross product category attribute extraction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3262–3270.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[17] Han Liu, Feng Zhang, Xiaotong Zhang, Siyang Zhao, Junjie Sun, Hong Yu, and Xianchao Zhang. 2022. Label-enhanced Prototypical Network with Contrastive Learning for Multi-label Few-shot Aspect Category Detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1079–1087.

[18] Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022. A Simple yet Effective Relation Information Guided Approach for Few-Shot Relation Extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 757–763. https://doi.org/10.18653/v1/2022.findings-acl.62

[19] Duangmanee (Pew) Putthividhya and Junling Hu. 2011. Bootstrapped Named Entity Recognition for Product Attribute Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Edinburgh, United Kingdom) *(EMNLP '11)*. Association for Computational Linguistics, USA, 1557–1567.

[20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[21] Martin Rezk, Laura Alonso Alemany, Lasguido Nio, and Ted Zhang. 2019. Accurate Product Attribute Extraction on the Field. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. 1862–1873. https://doi.org/10.1109/ICDE.2019.00202

[22] Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2018. NIH Public Access, 3132.

[23] Keiji Shinzato and Satoshi Sekine. 2013. Unsupervised Extraction of Attributes and Their Values from Product Description. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, 1339–1347. https://aclanthology.org/I13-1190

[24] Keiji Shinzato, Naoki Yoshinaga, Yandi Xia, and Wei-Te Chen. 2022. Simple and Effective Knowledge-Driven Query Expansion for QA-Based Product Attribute Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 227–234.

[25] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. 2022. Meta-learning for multi-label few-shot classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3951–3960.

[26] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30 (2017).

[27] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf

[28] 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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 47–55.

[29] Qifan Wang, Li Yang, Jingang Wang, Jitin Krishnan, Bo Dai, Sinong Wang, Zenglin Xu, Madian Khabsa, and Hao Ma. 2022. SMARTAVE: Structured Multimodal Transformer for Product Attribute Value Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 263–276.

[30] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Comput. Surv.* 53, 3, Article 63 (jun 2020), 34 pages. https://doi.org/10.1145/3386252

[31] J. D. Williams, A. Raux, D. Ramachandran, and A. Black. 2012. Dialog state tracking challenge handbook.. In *Technical report, Technical report, Microsoft Research*.

[32] Yuk Wah Wong, Dominic Widdows, Tom Lokovic, and Kamal Nigam. 2009. Scalable Attribute-Value Extraction from Semi-Structured Text. In *ICDM Workshop on Large-scale Data Mining: Theory and Applications*. http://www.computer.org/portal/web/csdl/doi/10.1109/ICDMW.2009.81

[33] Huimin Xu, Wenting Wang, Xinnian 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*. 5214–5223.

[34] 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)*. Association for Computational Linguistics, Online, 4694–4705. https://doi.org/10.18653/v1/2021.acl-long.362

[35] Kun Yan, Chenbin Zhang, Jun Hou, Ping Wang, Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. 2022. Inferring prototypes for multi-label few-shot image classification with word vector guided attention. (2022).

[36] 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 *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM*

'22). Association for Computing Machinery, New York, NY, USA, 1256–1265. https://doi.org/10.1145/3488560.3498377

[37] Zhuo Yang, Yufei Han, Guoxian Yu, Qiang Yang, and Xiangliang Zhang. 2019. Prototypical networks for multi-label learning. *arXiv preprint arXiv:1911.07203* (2019).

[38] Xinyang Zhang, Chenwei Zhang, Xian Li, Xin Luna Dong, Jingbo Shang, Christos Faloutsos, and Jiawei Han. 2022. OA-Mine: Open-World Attribute Mining for E-Commerce Products with Weak Supervision. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 3153–3161. https://doi.org/10.1145/3485447.3512035

[39] Fei Zhao, Yuchen Shen, Zhen Wu, and Xinyu Dai. 2022. Label-Driven Denoising Framework for Multi-Label Few-Shot Aspect Category Detection. *arXiv preprint arXiv:2210.04220* (2022).

[40] 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 and Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 1049–1058. https://doi.org/10.1145/3219819.3219839

[41] Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal Joint Attribute Prediction and Value Extraction for E-commerce Product. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 2129–2139. https://doi.org/10.18653/v1/2020.emnlp-main.166