What Matters for Shoppers: Investigating Key Attributes for Online Product Comparison

Nikhita Vedula, Marcus Collins, Eugene Agichtein, and Oleg Rokhlenko

Amazon, Seattle, USA
{veduln, collmr, eugeneag, olegro}@amazon.com

Abstract. Before making high-consideration purchase decisions, shoppers generally need to identify and evaluate products' key differentiating features or attributes. Many customers, however, lack the knowledge required to do so for all product domains. In this work, we investigate and analyze alternatives for identifying important product attributes, which customers can then use to compare candidate products. We propose an unsupervised attribute-ranking approach ReBARC, that combines both objective data from structured product catalogs, and subjective information from unstructured customer reviews, to suggest to the shopper the most important attributes to consider. Our detailed analysis of product attribute importance across various domains on a shopping website shows that ReBARC significantly outperforms prior efforts judged by both automated and human evaluation metrics. We also analyze the correlation and overlap between key product attributes detected by ReBARC, and those visible to customers during online product search.

1 Introduction and Background

E-commerce web sites contain a wealth of information describing the products they sell in the form of product features or attributes, which is largely factual and objective, and which is organized in a structured catalog. For instance, commonly available attributes for laptops include *brand*, *screen dimensions* and *memory*. These catalog attributes describe the product characteristics and help customers find and evaluate products for purchase. However, each product may have a large number of attributes, not all of which are equally useful. Searchers may not know in advance which attributes or product features are *important* to evaluate a given product. An important resource for customers to learn what attributes or features are most important is the opinion of other customers, in the form of *reviews* [31, 14, 6, 8]. Often, these reviews are quite detailed, and cover multiple product characteristics, attributes, and features useful to the review author. It is not feasible for customers to read a large set of reviews and aggregate multiple opinions to identify key product attributes.

A method to detect and suggest to the searcher the most important attributes can significantly improve their search and shopping experience in several ways. It would guide manufacturers to better help customers by choosing which attributes to highlight in product titles and descriptions. Important attributes could be

used as hints to help customers navigate retail websites or refine searches; or to offer appropriate product recommendations. For many customers, identifying these key attributes will educate them about the key considerations for the product category, and guide their comparison of multiple similar products, in product categories they are not yet familiar with (*e.g.*, Electronics). Such a method requires a high-quality, complete set of attributes, a way to rank them, and a source of data from which to compute the ranking. Several attempts have been made to extract product attribute names and their values from web pages using rule-based techniques [19, 27], naive Bayes and EM-based algorithms [11, 28], co-training [36], external dictionaries [29], feature engineering [17, 22, 24], active learning [39] and aspect extraction [7, 37, 26, 25, 10]. These methods do not generalize well across domains, and the expense of procuring manually labeled data makes them infeasible to be used at the scale of e-commerce. Distant supervision using general-purpose, open-source knowledge bases have been proposed to alleviate this cost [12, 38]. But they are limited by the accuracy and completeness of the external sources, to tackle which additional efforts would have to be made [15, 35, 33]. Unsupervised extraction of popular attributes or aspects from review text has been studied before [1, 12, 13, 34, 21]. But directly using keywords mentioned in customer reviews as aspects or attributes also often leads to a lower domain coverage and noisy, incoherent and redundant aspects which need further manual clean-up to avoid downstream errors.

We propose an approach, ReBARC (Review Based Attribute Ranker for Product Comparison) that ranks objective product catalog attributes and subjective product aspects, based on their presence in customer reviews and the sentiment of review authors towards these attributes. We use catalog data provided by product manufacturers, which is likely to be accurate and complete. ReBARC is *domain-agnostic* and *unsupervised*, requiring no manually labeled training data. ReBARC also avoids direct use of noisy review data, by mapping attribute mentions in reviews back to the more reliable structured catalog data.

2 ReBARC: Review-Based Attribute Ranking for Product Comparison

Data Collection: We utilize the Amazon Product Reviews (APR) [23] data as our primary source to develop and evaluate ReBARC. APR consists of a set of products, their associated customer reviews, and some metadata for each product, (i.e. product categories, similar products, and catalog attributes). We consider all reviews that other users have marked helpful at least once. Available catalog attributes include both generic attributes common across product categories such as *price, item weight, etc.*, and product- or category-specific attributes such as *screen size*. We also performed a web crawl on product pages from `www.amazon.com` to obtain the names of specific aspects or features separately rated by customers who bought a given product, and added these to our set of potential product attributes. We manually removed any attributes that are unlikely to influence users' buying choices (e.g. *date first available*). We ex-

periment on more than 10,000 unique products with an average of 64 attributes and 116,700 reviews per product category, as shown in Table 1. We now present our proposed method ReBARC, which involves ranking attributes based on their presence in reviews, as well as customer sentiment towards the attributes.

Popularity based Attribute Ranking: We use a light-weight unsupervised key-phrase extraction method based on text statistical features, YAKE [4], to extract a set of useful terms from the reviews linked to each product. We then segment each review into sentences, and extract only those sentences that either contain a useful term or a product attribute. This yields R useful review sentences per product. Since the product title is likely to contain useful attribute information identified by its sellers or manufacturers, we also append the product title to a sample of the R review sentences. We then use the pre-trained Sentence-BERT [30] model to compute embeddings for these sentences. We also compute Sentence-BERT embeddings for each attribute associated with these products. For each review sentence t in R, we select the top 3 catalog attributes for the product, using Maximal Marginal Relevance (MMR) [5] to rank the attributes for each product, based on their cosine similarity with the transformed review sentences:

$$MMR = \arg\max_{a_i \in A-S} [\lambda(sim(a_i, t) - (1-\lambda) \max_{a_j \in S} sim(a_i, a_j)]$$

where $a_i \in A$ and $a_j \in A$ denote attributes being ranked. S denotes the subset of attributes already selected for ranking. λ trades off between the similarity of the ranked attributes to the transformed review sentences and to each other. This ensures that highly ranked attributes are similar to both the review sentences and product titles, and are also diverse from each other to avoid redundancy. Of the resulting $3R$ attribute-sentence pairs, we pick the k most frequent attributes as the highest ranked attributes based on review popularity, for the given product.

Opinion based Attribute Re-Ranking: From Section 2, we obtain a list of k highly ranked attributes associated with each product, and also the review sentences that are relevant to each attribute. This approach considers both the seller-identified attributes with respect to the product title and catalog, as well as specific aspect features rated by customers, but does not yet consider customer *sentiment* with respect to the attributes mentioned in reviews. To perform a secondary ranking using sentiment, we utilize a RoBERTa [20] model fine-tuned on the SST-2 sentiment detection benchmark corpus [32]. This model outputs a sentiment score for each review sentence relevant to an attribute, i.e. how positive or negative the sentence is. We assume that the sentiment score of a sentence relevant to attribute a represents the sentiment towards a itself. We then average the absolute sentiment scores for each attribute a over all sentences linked with a. We use the absolute value because we want to find attributes customers feel strongly about, whether they feel negatively or positively. Finally, we re-rank the list of attributes obtained earlier using the aggregated sentiment scores, yielding the final ranked list of *top-k* attributes for each product. Therefore, ReBARC ensures that attribute importance is evaluated based on direct and indirect mentions of product attributes by buyers in their reviews; as well as the (positive

or negative) opinions of customers towards these attributes. Highly-ranked, key attribute values for similar products can then be compared by users to make purchase decisions. We can improve our technique of finding the sentiment of a review sentence by extracting sub-sentence fragments, and aggregating the sentiment score of each fragment as the sentiment score of the sentence. There are also unusual ways of mentioning certain attributes that might be missed by our sentence embedding technique, which might need to be solved by manual intervention or supervision. We leave these directions for future work.

3 Experiments and Results

Experimental Setup: Our first baseline (*SRA*) consists of those attributes customers used to refine or filter product searches on an online shopping website which most frequently lead to customers purchases or adding products to their shopping cart. Our second baseline (*QAC*) consists of the top attributes identified from query auto-completion [3] logs of the same online shopping website, that assist in automatically completing customers' search queries for a product. The above two approaches are state-of-the-art, optimal indicators of attribute importance based on real-world search and purchase behavior of millions of customers using this shopping website. All data was aggregated, anonymized and limited to targeted and relevant information (product names, attribute names and values), to protect customer privacy. We also compare ReBARC with a recent high-performing unsupervised aspect extraction technique, CAt [34]. We evaluated other existing techniques [1, 12], using TF-IDF weighting of attributes extracted from reviews, and unsupervised methods using CRFs [18] to extract attributes from review text. These methods did not outperform any of our other baselines, so to save space we omit their results in Table 2. Since ReBARC is completely unsupervised, we do not compare it with any supervised methods.

We performed multiple crowd-sourced user studies to assess the performance of ReBARC, and followed recommended practices [2] to ensure good quality output from crowd workers. For a given product, we presented to annotators one randomly chosen attribute from the top 5 important attributes identified by our model, and asked the annotators if they would consider that attribute important if purchasing that product (Table 1, inter-annotator agreement Cohen's $\kappa = 0.77$). We also combined and shuffled the top 5 attributes each from ReBARC and the baselines for specific products, and presented a list of about 15 unique attributes to crowd workers. We asked them to pick the top 3 and top 5 attributes that they thought would help them the most in buying that particular product, or in comparing other similar product options of the same type (Table 2, Cohen's $\kappa = 0.65$, which indicates a substantial inter-annotator agreement [9]). We manually inspected and cleaned the task to ensure that crowd workers were not asked to judge attributes that required any specialized domain expertise. All parameters of ReBARC were tuned based on performance on a validation set, which we created based on the above ground truth human annotations.

Experimental Results: About 54% of the important attributes ranked highly

Table 1: Key attributes detected by ReBARC and chosen as important by annotators. 'Num.' and 'cat.' denote numerical and categorical valued attributes.

Product Category (#Products, #Attributes, #Reviews)	**Human imp. attrs.**	**Human imp. num. attrs.**	**Human imp. cat. attrs.**	**Sample key attributes frequently detected by ReBARC per category**
Home (2319, 61, 150K)	0.66	0.78	0.55	*color, assembly, easyToClean*
Electronics (3267, 84, 339K)	0.71	0.82	0.56	*price, display, color, resolution*
Tools (1218,76,291K)	0.73	0.82	0.55	*durability, easy to install rating*
Beauty (546, 48, 66K)	0.77	0.82	0.71	*brand, skinType, valueForMoney*
Appliances (1104, 78, 91K)	0.81	0.84	0.72	*batteries, price, brand, rating*
Avg (all 10 categories)	0.71	0.77	0.61	N/A

Table 2: Evaluating the top k ranked important attributes using human evaluation and the metrics MAP@k and NDCG@k, for $k = 3$ and 5. Best performances are in bold. R, S, Q and C denote ReBARC, SRA, QAC, and CAt respectively.

Product Category	**MAP@5**				**MAP@3**				**NDCG@5**				**NDCG@3**			
	R	**S**	**Q**	**C**	**R**	**S**	**Q**	**C**	**R**	**S**	**Q**	**C**	**R**	**S**	**Q**	**C**
Home	**0.51**	0.38	0.34	0.42	**0.32**	0.24	0.19	0.26	**0.57**	0.36	0.12	0.45	**0.43**	0.25	0.08	0.36
Electronics	**0.5**	0.35	0.36	0.4	**0.4**	0.2	0.24	0.26	**0.48**	0.27	0.14	0.39	**0.45**	0.13	0.05	0.34
Tools	**0.5**	0.36	0.35	0.39	**0.3**	0.21	0.18	0.21	**0.55**	0.32	0.18	0.44	**0.44**	0.19	0.1	0.34
Pets	**0.52**	0.37	0.34	0.41	**0.42**	0.33	0.25	0.31	**0.6**	0.36	0.17	0.5	**0.48**	0.34	0.1	0.37
Beauty	**0.6**	0.32	0.32	0.43	**0.35**	0.14	0.15	0.22	**0.65**	0.13	0.12	0.5	**0.52**	0.1	0.05	0.41
Grocery	**0.6**	0.33	0.35	0.46	**0.5**	0.22	0.23	0.37	**0.68**	0.11	0.18	0.51	**0.6**	0.1	0.13	0.47
Appliances	**0.57**	0.37	0.35	0.41	**0.48**	0.28	0.21	0.33	**0.7**	0.28	0.19	0.57	**0.64**	0.17	0.1	0.5

by ReBARC were numerically-valued. Table 1 shows that annotators chose 71% of our key product attributes as useful for making purchase decisions. We observe that more than 60% of attributes available as search refinement filters or recommended during query auto-completion are categorical-valued. On the contrary, ReBARC detects a good mix of categorical and numerical valued key attributes, across different product groups. Overall, annotators preferred numerically- over categorically-valued attributes. Customers are thus likely to benefit from access to more numerically-valued attributes during their product search and comparison process. Table 2 evaluates ReBARC and three baselines in ranking important product attributes. CAt [34] outperforms SRA and QAC for most product categories. ReBARC significantly outperforms all baselines by 10-20% across product categories, as per Mean Average Precision [40] and Normalized Discounted Cumulative Gain [16]. Inspecting a random sample of products also showed that key attributes detected by ReBARC are diverse with less repetition.

Discussion: Our results show that more than 70% of the review sentences we analyzed either explicitly mention the names of attributes (58% of the time), or have a high cosine similarity > 0.7 to a catalog attribute (42% of the time). A wide range of numerical and categorical attributes identified by ReBARC were found useful by our human annotators. Most prior work extracts attribute names directly from review text, which is further used to identify key product attributes. This can cause ambiguity and redundancy in the important attributes detected, since the same catalog attribute can be referred to by different names in reviews (e.g. for a laptop, both *performance* and *speed* refer to a single *pro-*

cessor attribute). In contrast, ReBARC links customer opinions taken from user reviews to existing product attributes identified by product retailers or manufacturers. Thus, it maintains consistency in the detected key attributes and avoids ambiguity and redundancy despite being completely unsupervised. Interestingly, sentiment-based attribute re-ranking improves performance for specific product categories only. For instance, in *Electronics*, nearly 60% of the attributes are discussed in a neutral, descriptive, rather than opinionated, way. Some attributes are also more frequently referred to than others in reviews (e.g. *network speed* vs *frequency band* for routers). In these cases, signals from reviews could be combined with search-based popularity for additional improvements.

Our evaluation reveals that the overlap between important attributes detected by ReBARC, and those sourced by search refinements or query auto-completions, is lower than 50% across various product categories. The search logs of the shopping website under consideration show that a large fraction of the search filters and query auto-completions suggest generic attributes (e.g. *price, brand, delivery speed*). In contrast, our model identifies both generic and product-specific attributes. Annotators perceived a product-specific attribute as more useful than a generic attribute for product comparison in more than 65% cases. For instance, ReBARC identified *wireless network speed* as a popular and important attribute based on reviews for routers. However, the e-commerce engine does not suggest anything related to "network speed" as a filter or auto-completion suggestion when searching for any of the diverse queries *'router', 'router wifi', 'router speed', 'router internet', 'router wireless', or 'router network'*. Incorporating attributes identified from reviews into the search interface could improve the search and shopping experience, especially for more technical product categories such as *Electronics* or *Tools and Home Improvement*. Understanding the meaning or values of catalog attributes for certain product categories may require the searcher to possess domain knowledge. Such attributes could be referred to by more common, easier to understand terminology from reviews, captured by ReBARC. For example, the term *image quality* from reviews can refer to more technical attributes such as *refresh rate or resolution*. Thus, our insights imply that automatic product comparison and customer education would benefit from a diverse set of both generic and product-specific attributes.

4 Conclusions

We presented an unsupervised approach, ReBARC, that uses data from structured product catalogs and customer opinions from reviews to automatically identify key product features useful for online shopping and product comparison. ReBARC significantly outperforms strong baselines on diverse metrics and product domains. We also studied the correlation between product attributes of interest to customers based on reviews, and those available to them for search on shopping websites. In future, we plan to actively use customer behavior and shopping history for detecting key attributes, and personalizing attribute ranking for customers.

Bibliography

[1] Bing, L., Wong, T., Lam, W.: Unsupervised extraction of popular product attributes from e-commerce web sites by considering customer reviews. ACM TOIT (2016)

[2] Buhrmester, M., Kwang, T., Gosling, S.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? (2016)

[3] Cai, F., de Rijke, M.: A survey of query auto completion in information retrieval. Foundations and Trends in Information Retrieval (2016)

[4] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., Jatowt, A.: Yake! keyword extraction from single documents using multiple local features. Inf. Sci. (2020)

[5] Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: ACM SIGIR (1998)

[6] Carmel, D., Lewin-Eytan, L., Maarek, Y.: Product question answering using customer generated content-research challenges. In: ACM SIGIR (2018)

[7] Chen, G., Tian, Y., Song, Y.: Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In: COLING (2020)

[8] Chen, S., Li, C., Ji, F., Zhou, W., Chen, H.: Driven answer generation for product-related questions in e-commerce. In: ACM WSDM (2019)

[9] Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)

[10] Da'u, A., Salim, N.: Aspect extraction on user textual reviews using multi-channel convolutional neural network. PeerJ Computer Science **5** (2019)

[11] Ghani, R., Probst, K., Liu, Y., Krema, M., Fano, A.: Text mining for product attribute extraction. ACM SIGKDD Explorations Newsletter (2006)

[12] Giannakopoulos, A., Musat, C., Hossmann, A., Baeriswyl, M.: Unsupervised aspect term extraction with b-lstm & crf using automatically labelled datasets. arXiv preprint arXiv:1709.05094 (2017)

[13] He, R., Lee, W.S., Ng, H.T., Dahlmeier, D.: An unsupervised neural attention model for aspect extraction. In: ACL (2017)

[14] Hirschmeier, S., Egger, M.: Social product search–enhancing product search with mined (sparse) product features (2018)

[15] Huynh, V.P., Papotti, P.: A benchmark for fact checking algorithms built on knowledge bases. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 689–698 (2019)

[16] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. TOIS (2002)

[17] Kozareva, Z., Li, Q., Zhai, K., Guo, W.: Recognizing salient entities in shopping queries. In: ACL (2016)

[18] Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)

[19] Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: WWW (2005)

[20] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
[21] Luo, L., Ao, X., Song, Y., Li, J., Yang, X., He, Q., Yu, D.: Unsupervised neural aspect extraction with sememes. In: IJCAI (2019)
[22] More, A.: Attribute extraction from product titles in ecommerce. arXiv preprint arXiv:1608.04670 (2016)
[23] Ni, J., Li, J., McAuley, J.: Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: EMNLP-IJCNLP (2019)
[24] Petrovski, P., Bizer, C.: Extracting attribute-value pairs from product specifications on the web. In: International Conference on Web Intelligence (2017)
[25] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M., Eryiğit, G.: SemEval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (2016)
[26] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 task 4: Aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (2014)
[27] Popescu, A., Etzioni, O.: Extracting product features and opinions from reviews. In: Natural language processing and text mining (2007)
[28] Probst, K., Ghani, R., Krema, M.and Fano, A., Liu, Y.: Semi-supervised learning of attribute-value pairs from product descriptions. In: IJCAI (2007)
[29] Putthividhya, D., Hu, J.: Bootstrapped named entity recognition for product attribute extraction. In: EMNLP (2011)
[30] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: EMNLP-IJCNLP (2019)
[31] Retail, T.: They say they want a revolution - price water house (2016), https://www.pwc.es/es/publicaciones/retail-y-consumo/assets/total-retail-2016.pdf
[32] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: EMNLP (2013)
[33] Thorne, J., Vlachos, A.: Evidence-based factual error correction. arXiv preprint arXiv:2106.01072 (2021)
[34] Tulkens, S., van Cranenburgh, A.: Embarrassingly simple unsupervised aspect extraction. ArXiv **abs/2004.13580** (2020)
[35] Vedula, N., Parthasarathy, S.: Face-keg: Fact checking explained using knowledge graphs. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. pp. 526–534 (2021)
[36] Wu, B., Cheng, X., Wang, Y., Guo, Y., Song, L.: Simultaneous product attribute name and value extraction from web pages. In: ACM WI-IAT (2009)

[37] Xu, H., Liu, B., Shu, L., Yu, P.S.: Double embeddings and cnn-based sequence labeling for aspect extraction. ArXiv **abs/1805.04601** (2018)
[38] Yang, Y., Chen, W., Li, Z., He, Z., Zhang, M.: Distantly supervised ner with partial annotation learning and reinforcement learning. In: International Conference on Computational Linguistics (2018)
[39] Zheng, G., Mukherjee, S., Dong, X., Li, F.: Opentag: Open attribute value extraction from product profiles. In: ACM SIGKDD (2018)
[40] Zhu, M.: Recall, precision and average precision. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (2004)