E-commerce Query Classification Using Product Taxonomy Mapping: A Transfer Learning Approach

Michael Skinner
research@mcskinner.com

Surya Kallumadi
The Home Depot
surya@ksu.edu

ABSTRACT

In web search, query classification (QC) is used to map a query to a user's search intent. In the e-commerce domain, user's product search queries can be broadly categorised into product specific queries and category specific queries [9]. In these instances, accurate classification of queries will help with identifying the right product categories from which relevant products can be retrieved. Thus, mapping a query to a pre-defined product taxonomy is an important step in e-commerce query understanding pipeline. A typical e-commerce website has thousands of categories, and curating a labeled data set for query classification is expensive, time consuming, and labor intensive. In addition, product search queries are short, and the vocabulary changes over time as the catalogue evolves. Reducing this effort of generating query-category labels would save time and resources. In this work we show how an existing product-taxonomy mapping can improve query classification, and reduce the need for labeled data, using transfer learning. Our results demonstrate that such an approach can match, and often exceed, the performance of direct training with a smaller computational budget. We further explore how performance varies as the amount of available training data varies, and show that transfer learning is most useful when the target data set size is small. In addition, we make available a large query data set of 535, 506 unique e-commerce labeled queries, mapped over 58 categories. The results and transfer learning approaches presented in this work can act as strong baselines for this collection and task.

CCS CONCEPTS

• **Information systems → Clustering and classification**;

KEYWORDS

Test Collection, e-Commerce, Query Classification

1 INTRODUCTION

In the e-commerce domain query understanding can have a significant impact on user satisfaction. An incorrectly interpreted query can lead to search abandonment by the user, resulting in lower conversion rates. E-commerce queries are usually short and lack linguistic structure, and they can be ambiguous as a result. For example the query `'battery lawn tractor'`, can be interpreted as `'battery for lawn tractor'` or `'battery operated lawn tractor'`.

In product search, the objective of query classification is to map a user query to a pre-defined product category. QC can improve the relevance of search results while preserving the recall. A typical e-commerce site such as `Amazon.com` can have millions of products, and thousands of product categories of various granularities. Curating a query-category labeled data set with good coverage over all the categories is expensive, labor intensive, and can take a long time. Approaches that can reduce the effort needed to categorize the search queries can significantly improve the performance of QC. In this work, we propose a transfer learning approach for QC by using product titles. As the products in the domain are mapped to a well defined product taxonomy, the product mapping can be exploited to improve QC, and reduce the need for labeled data.

Transfer learning has proven to be an effective technique to improve the performance of various tasks in computer vision and natural language processing (NLP) [1]. The goal of transfer learning is to utilize knowledge present within a *source* domain to improve a task within a *target* domain. Neural network and deep learning based transfer learning approaches have been shown to be quite useful to improve the performance of a wide range of target tasks in NLP [7]. To demonstrate transfer learning for QC in the e-commerce domain, we use `Amazon.com` titles as the source data set [5], and queries obtained by crawling `Amazon.com` auto-complete service as target data set.

Academic research for e-commerce query classification task has been limited because of a lack of availability of labeled data. Through this work, and the query-category data set made available, we hope to facilitate progress in this research area. In addition to the introduction of a new data set, our contributions are as follows: 1) We present a methodology for this domain-specific transfer learning, in which the source model is tuned as a classifier on a similar problem. 2) We demonstrate that such an approach can be leveraged to speed training and improve results when compared to direct training. 3) We explore the impact of target data size on both direct and transferred models, showing that transfer learning improves more on direct training as the target training data shrinks.

2 RELATED WORK

In the query classification challenge, organized by ACM KDD cup 2005 competition, the task was to categorize 800, 000 web queries into 67 predefined categories [3]. The data set for this challenge contained 111 queries with category mappings, and the queries in the test data set can be tagged by up to 5 categories. The submissions were evaluated on an 800 query subset of the complete data set. This competition highlighted the challenge of assigning labels to queries.

In: J. Degenhardt, S. Kallumadi, U. Porwal, A. Trotman (eds.):
Proceedings of the SIGIR 2019 eCom workshop, July 2019, Paris, France, published at http://ceur-ws.org

Product Titles	**Category**
Compaq 256MB 168-Pin 100Mhz DIMM SDRAM for Compaq Proliant	Electronics
EK Ekcessories 10708C-BLUE-AM Blue Jeep Visor Clip	Automotive
NHL Chicago Blackhawks Franchise Fitted Hat, Black, Extra Large	Sports & Outdoors
Sesame Street Robe with Embroidered Washcloth	Health & Personal Care
Emerica Men's The Westgate Skate Shoe	Clothing, Shoes & Jewelry
Queries	**Category**
13mm wrench	tools
hip action zukes peanut butter	pets
nerf guns under 30 dollars	toys-and-games
bernaise sauce mix	grocery
door lever lock child proof	baby-products

Table 1: Examples from the source (titles) and target (queries) data sets.

Lin et al. propose using implicit feedback from user clicks as a signal to collect training data for QC in e-commerce domain [4]. We consider this work to be complementary to the transfer learning approach we propose in this paper. Leveraging user click stream data and the product hierarchy together can be used to improve the overall system performance. Click stream data is useful when a sufficient amount of user behavior has been observed for a category, but this fails for new categories and items. The transfer learning approach exploiting product titles does not suffer from item and category *cold start*.

Sondhi et al. identify a taxonomy of e-commerce queries intents, based on search logs and user behavior data [9]. This work identifies five categories of e-commerce queries based on user search behavior: 1) Shallow Exploration Queries, 2) Targeted Purchase Queries, 3) Major-Item Shopping Queries, 4) Minor-Item Shopping Queries, and 5) Hard-Choice Shopping Queries. This paper highlights the complexity of user intent in the e-commerce domain, and proposes techniques for leveraging these insights.

3 DATA COLLECTION AND DATA SET

Domain adaptation and transfer learning usually requires two data sets, a source data set and a target data set. For supervised tasks such as QC, transfer learning would help in scenarios where we have very little training data in the target data set, and lots of data in the source data set. Also, the source and target data set should have similar characteristics. In this work, as product titles and queries share a similar vocabulary, we chose product titles as the source data set. McAuley et al. [5] provide a crawl of `Amazon.com`'s product pages including 142.8 million reviews, 9.43 million products, and 6.83 million titles[1]. We utilize the titles data available in this data set as the source data for transfer learning.

As no product-query data sets are publicly available for QC, we leveraged `Amazon.com`'s auto-completion to generate e-commerce queries[2]. In addition to providing suggestions for partial queries, auto-complete also provides high level candidate categories for suggested queries. These query-category results serve as our target data set for the QC task. The seeds for *auto-complete crawl* were common terms and phrases found in the data set by McAuley et al. In addition, we used random alpha-numeric character combination as seeds for the query crawl. A total of $535,506$ query-category labels were obtained by this exercise. To ascertain the accuracy of this data, we manually evaluated 1000 randomly sampled queries from this data set. The query-category labels suggested by auto-complete had an accuracy of 98.6%. The auto-complete crawl was performed over a duration of 1 week, in December 2018. The queries in the resulting data-set were mapped to 58 high level categories.

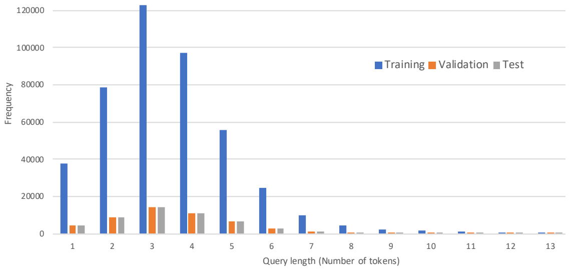

Figure 1: Query token-length distributions across target data splits.

	Source	**Target Data - Queries**		
	(Titles)	**Train**	**Val.**	**Test**
Documents	$6,835,398$	$435,506$	$50,000$	$50,000$
Num. bytes	$1-2000$	$1-69$	$3-69$	$2-69$
Avg num. bytes	59.98	21.32	21.34	21.37
Token length	$1-434$	$1-14$	$1-14$	$1-14$
Avg token length	9.45	3.52	3.52	3.52

Table 2: Statistics of each part of the source and target data.

3.1 Data Splits

Both the source and target data sets are split into training, validation, and test sets, stratified by category. This resulted in $5,811,656$ training examples for the source data, $500,000$ validation examples and $500,000$ test examples. The target data had $435,506$ training examples, with $50,000$ examples reserved for validation and test sets each. The target training data was also progressively sub-sampled to create smaller training sets of 50%, 20%, and 10% of the original data, each a subset of the previous sample. In Figure 1 we can see that the length of queries is similarly distributed across the 3 splits. Both the validation set and the training set show a Pearson's correlation of > 0.99 with the test set. Due to the use of stratified sampling, the category distributions over the three sets are similar.

3.2 Data Characteristics

Table 1 shows examples of product titles and queries from the source and target data set, respectively. Table 2 shows the high level statistics of the source and target data sets. While the average length of a title is 9.45, queries are much shorter (3.52 tokens). This significant difference in query and title lengths poses an interesting transfer learning challenge.

4 SYSTEM ARCHITECTURE DESCRIPTION

Recent work in NLP has shown the wide utility of Long Short-Term Memory (LSTM) architectures for transfer learning tasks [6]. Howard and Ruder used a pre-trained LSTM architecture to achieve state-of-the-art results on several text classification tasks [2]. The Balanced Pooling View (BPV) architecture, which builds on these

[1] http://jmcauley.ucsd.edu/data/amazon/

[2] http://completion.amazon.com/api/2017/suggestions

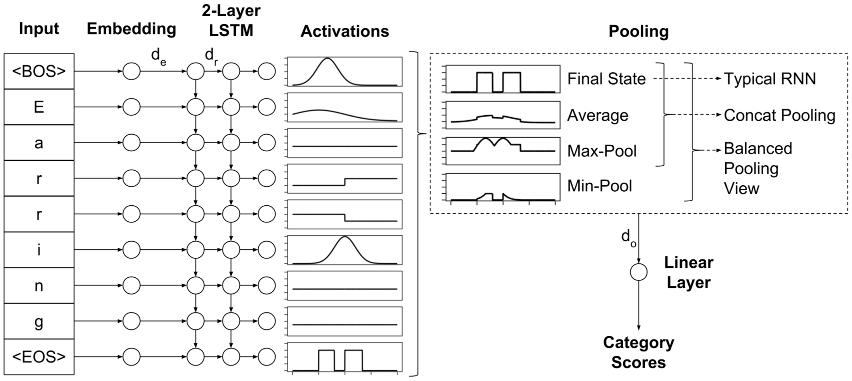

Figure 2: An illustration of the BPV architecture.

approaches, has been shown to be effective for product taxonomy classification tasks [8].

The model architecture, which can be seen in Figure 2, is centered around a character-level LSTM, which is fed via an a embedding. The time series output from the Recurrent Neural Network (RNN) is then summarized in 4 ways: by taking the last value as in a typical RNN architecture, and then with mean-pooling, max-pooling, and min-pooling. Those 4 summaries are concatenated and fed through a linear layer with output size equal to the number of categories. When transferring, only the output layer needs to be replaced, in order to accommodate the new category space. The embedding size, RNN width and depth, and dropout settings are all set as in [8].

On the target problem, we explore two different training styles, 1) target only direct training and 2) transfer learning from a source model. Direct training only uses the target data, without reference to either the source model or the source data. Transfer learning uses the source model to initialize network weights, replacing the output layer to accommodate the new category set, and then otherwise proceeding as before. *Adam optimization* was found to be consistently better than *stochastic gradient descent* (SGD) and is used for all target models. Cross-entropy loss is used throughout.

Final hyper-parameters were tuned using a grid search around those initial values, varying the learning rate schedule and peak learning rate, as well as the number of training epochs for direct training. Transfer learning was fixed at 5 epochs throughout, since any increase in the number of epochs led to overfitting and an increasing validation loss. This process was performed separately for direct training and transfer learning, as well as for each of the 4 data scales.

Hyper-parameters with consistently strong validation results were then chosen for each of the two training styles. A learning rate of 0.003 was best for all variants. A linearly decreasing "burndown" schedule was better than `1cycle` or a flat learning rate for transfer. Direct training was most effective with 10 epochs when trained on subsets of the target data, but better still with 20 epochs on a full 100% of the target data. Once settled, these parameters were used in 4 independent training runs for each training style and data scale. Each model was used to make predictions over the test set, and the results are based on these predictions.

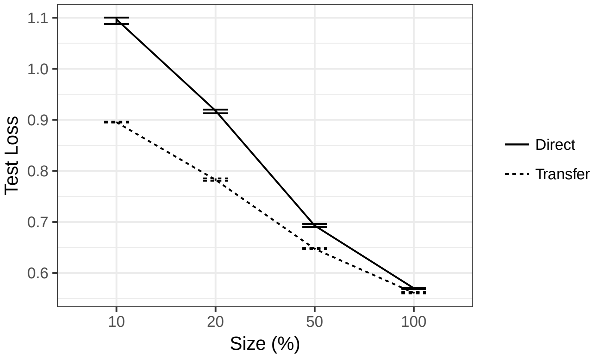

Figure 3: Cross-entropy loss on the test set, with varying target data size.

5 EVALUATION

We report cross-entropy loss, accuracy, precision, recall, and $F1$ scores for our models. As the queries are not uniformly distributed across the categories, we use weighted precision, recall, and $F1$ to measure the performance of the approaches on the test data. If (P_i), (R_i) and $(F1_i)$ are precision, recall, and $F1$ scores for each category c_i, then the corresponding weighted metrics can be calculated as:

$$P_w = \sum_{i=1}^{K} \frac{n_i}{N} P_i \qquad (1)$$

$$R_w = \sum_{i=1}^{K} \frac{n_i}{N} R_i \qquad (2)$$

$$F1_w = \sum_{i=1}^{K} \frac{n_i}{N} F_{1_i} \qquad (3)$$

6 RESULTS

Figure 3 shows the results for test loss as the amount of target data varies, for each of the two training approaches. The advantages of transfer learning are most apparent at low data scales, where it produces significantly better results. The two approaches eventually converge in performance as target data becomes fully available. Figure 4 shows the equivalent results for accuracy. In this case the performance difference is not as large, and direct training closes the gap at 50% of the target data. This corresponds to a regime in which the training loss continues to drop rapidly while validation loss levels off, which might indicate overfitting.

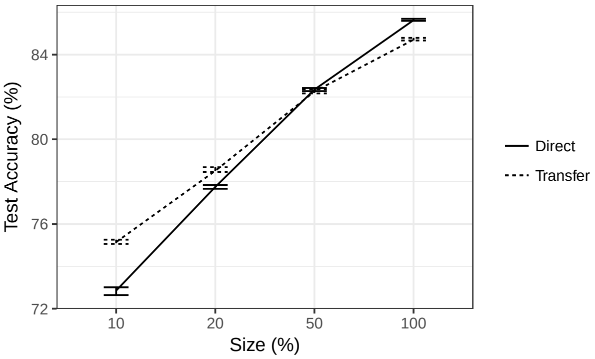

Figure 4: Accuracy on the test set for various data scales.

Target Size	Source + Target			Target only		
	P_w	R_w	$F1_w$	P_w	R_w	$F1_w$
10%	0.757	0.757	0.754	0.733	0.734	0.732
20%	0.791	0.790	0.788	0.782	0.783	0.781
50%	0.828	0.828	0.826	0.828	0.829	0.827
100%	0.852	0.852	0.851	0.862	0.861	0.860

Table 3: Comparing the performance of transfer learning and direct target-only training.

Table 3 shows the overall weighted precision, recall, and $F1$ scores for each training variant across the different target data scales. Recall is equal to the accuracy metrics reported in Figure 4. Table 4 shows the per-category results in the case when the target training data set is small (10%), for categories with at least 100 test examples. Transfer learning is able to improve $F1$ for nearly all categories, sometimes significantly, and for categories that were both difficult as well as easy for the directly trained model. Transfer learning was particularly helpful for rare categories. The top 6 $F1$ improvements (bolded) were achieved on the 6 categories with the fewest examples in the 10% subset of target training data. This highlights the benefit of a transfer learning approach for cold start categories and items.

7 CONCLUSION

Our results show that product-title data is an effective pre-training source for query-taxonomy classification. When there is not much training data, transfer learning improves the quality of the final target models. Although the results converge for larger target data sets, we observe that pre-trained transfer learning models converge in fewer epochs than models trained only on the target data set.

This convergence is noteworthy and worth exploring in more detail. The implication is that, at a certain data scale, the source model does not contain any information that is more useful than that in the target data. One possible reason for this is that the model architecture can only encode so much information, and it may be the case that the full target data can saturate it. If so, then increasing the size of the pre-trained source model might lead to further improvements.

Category	Source + 10% Target			10% Target only		
	P	R	F1	P	R	F1
fash-wom-shoes	0.887	0.859	0.873	0.835	0.808	0.821
pets	0.903	0.820	0.860	0.882	0.813	0.846
mobile	0.860	0.827	0.844	0.862	0.823	0.842
fash-wom-cloth	0.847	0.837	0.842	0.792	0.753	0.772
beauty	0.835	0.830	0.833	0.802	0.787	0.795
garden	0.821	0.836	0.828	0.787	0.828	0.807
fash-wom-jlry	0.755	0.895	0.819	0.733	0.777	0.754
grocery	0.784	0.831	0.807	0.750	0.779	0.764
baby-products	0.824	0.727	0.772	0.773	0.655	0.709
electronics	0.712	0.844	0.772	0.741	0.809	0.774
automotive	0.705	0.776	0.739	0.686	0.761	0.721
toys-and-games	0.738	0.732	0.735	0.698	0.693	0.695
videogames	0.782	0.691	0.734	0.800	0.740	0.769
hpc	0.726	0.719	0.722	0.701	0.697	0.699
office-products	0.749	0.698	0.722	0.711	0.691	0.701
sports-&-fitness	0.719	0.661	0.689	0.685	0.616	0.649
arts-crafts	0.740	0.636	0.684	0.692	0.611	0.649
fash-mens-cloth	0.681	0.676	0.679	0.641	0.560	0.598
lawngarden	0.763	0.606	0.676	0.716	0.582	0.642
tools	0.678	0.670	0.674	0.651	0.668	0.660
fan-shop	0.745	0.597	0.663	0.631	0.430	0.511
mi	0.761	0.577	0.656	0.694	0.569	0.625
outdoor-rec	0.715	0.526	0.606	0.680	0.538	0.601
industrial	0.579	0.355	0.440	0.506	0.371	0.428
appliances	0.580	0.348	0.435	0.564	0.297	0.389

Table 4: Per-category results on 10% of the target data, for categories with at least 100 test examples.

In addition, we make available a large query-category labeled data set which can facilitate additional progress in this research area. This data provides scope for research tasks such as query intent mining, query segmentation and query scoping.

REFERENCES

[1] Hal Daumé, III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly Easy Semi-supervised Domain Adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP 2010)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 53–59. http://dl.acm.org/citation.cfm?id=1870526.1870534

[2] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. (2018). arXiv:arXiv:1801.06146

[3] Ying Li, Zijian Zheng, and Honghua (Kathy) Dai. 2005. KDD CUP-2005 Report: Facing a Great Challenge. *SIGKDD Explor. Newsl.* 7, 2 (Dec. 2005), 91–99. https://doi.org/10.1145/1117454.1117466

[4] Y. Lin, A. Datta, and G. D. Fabbrizio. 2018. E-commerce Product Query Classification Using Implicit UserâĂŹs Feedback from Clicks. In *2018 IEEE International Conference on Big Data (Big Data)*. 1955–1959. https://doi.org/10.1109/BigData.2018.8622008

[5] J McAuley, R Pandey, and J Leskovec. 2015. Inferring Networks of Substitutable and Complementary Products. In *KDD 2015*. 785–794.

[6] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. (2017). arXiv:arXiv:1708.02182

[7] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How Transferable are Neural Networks in NLP Applications?. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 479–489. https://doi.org/10.18653/v1/D16-1046

[8] M Skinner. 2018. Product Categorization with LSTMs and Balanced Pooling Views. In *Proceedings of the 2018 SIGIR Workshop On eCommerce*.

[9] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and Chengxiang Zhai. 2018. A taxonomy of queries for e-commerce search. In *41st International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR 2018. Association for Computing Machinery, Inc, 1245–1248. https://doi.org/10.1145/3209978.3210152