Transfer Learning for E-commerce Query Product Type Prediction

Anna Tigunova
Amazon
Berlin, Germany
tigunova@amazon.com

Thomas Ricatte
Amazon
Luxembourg, Luxembourg
tricatte@amazon.com

Ghadir Eraisha
Amazon
Luxembourg, Luxembourg
eraishag@amazon.com

Abstract

Getting a good understanding of the customer intent is essential in e-commerce search engines. In particular, associating the correct *product type* to a search query plays a vital role in surfacing correct products to the customers.

Query product type classification (Q2PT) is a particularly challenging task because search queries are short and ambiguous, the number of existing product categories is extremely large, spanning thousands of values. Moreover, international marketplaces face additional challenges, such as language and dialect diversity and cultural differences, influencing the interpretation of the query.

In this work we focus on Q2PT prediction in the global multi-locale e-commerce markets. The common approach of training Q2PT models for each locale separately shows significant performance drops in low-resource stores. Moreover, this method does not allow for a smooth expansion to a new country, requiring to collect the data and train a new locale-specific Q2PT model from scratch.

To tackle this, we propose to use *transfer learning* from the high-resource to the low-resource locales, to achieve global parity of Q2PT performance. We benchmark the per-locale Q2PT model against the *unified* one, which shares the training data and model structure across all worldwide stores. Additionally, we compare locale-aware and locale-agnostic Q2PT models, showing the task dependency on the country-specific traits.

We conduct extensive quantiative and qualitative analysis of Q2PT models on the large-scale e-commerce dataset across 20 worldwide locales, which shows that unified locale-aware Q2PT model has superior performance over the alternatives.

CCS Concepts

• **Information systems** → **Query intent**.

Keywords

Query Understanding, Product Search

ACM Reference Format:
Anna Tigunova, Thomas Ricatte, and Ghadir Eraisha. 2018. Transfer Learning for E-commerce Query Product Type Prediction. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Problem statement. Query understanding in e-commerce extracts customer shopping intent from their search queries, by classifying the query as having a target brand, color, size, etc. The extracted attributes are extremely important for search results ranking and filtering, query augmentation, recommendations and many other usecases [3]. One of the most critical components in query understanding isa Query-to-Product Type (Q2PT) classifier, which associates customer search query with a product type (PT) that the customer intended to browse.

Q2PT signal has a direct impact on the customer experience, as it can significantly alter search results shown to the user. For example, for a search query "*harry potter mug*", even if the matched Harry Potter *book* has a high TF-IDF score, it should not be surfaced in the search results, since the customer is interested in another product type (*mugs*). Moreover, Q2PT signal helps to optimize the computation of the search results, narrowing the retrieval to a specific product category shard [7], yielding search latency improvements.

Query category prediction has received significant attention in related works [6, 9, 17–19]: the proposed approaches target various challenges associated with Q2PT prediction, such as short queries [6, 9], long-tail queries [17, 18], product type hierarchy [19], etc. However, to the best of our knowledge, there has not yet been studies, which target the issues of Q2PT classification in the multi-locale setting.

In international marketplaces, query understanding and ranking models are often trained on a per-locale basis, using the data from a single store [1, 10]. This approach ensures that the peculiarities of various locales are captured: for instance, in case of Q2PT prediction, same keywords may convey different product type intents, depending on the store (query 'pants' in UK would mean that the user is searching for *underwear*, while in the US it means the intent to buy *trousers*); Bonab et al. [1] show that users in multiple locales have different preferences for the same product set.

Most related studies develop and test query understanding models for high-resource locales, such as United States, achieving remarkable results. However, many recently launched stores in new countries suffer from data shortage [5], which is a major blocker for applying per-locale Q2PT models.

The issue is further aggravated by the long-tail nature of the product category distribution; in large online stores there are thousands of product types, with a small fraction of most popular categories dominating the distribution [16]. This discrepancy is amplified in the emerging stores, as shown in Figure 1, where we use our experimental dataset to plot the number of queries per product type, sorted by their frequency, in the emerging locale and aggregated worldwide. In both graphs, we observe that the majority of queries

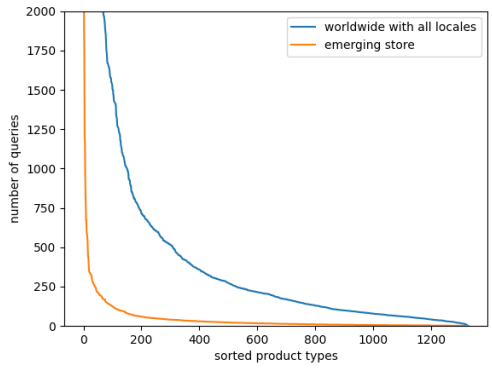


Figure 1: Query distribution for product types worldwide versus an emerging store on an experimental data sample including 20 locales and 1414 product types.

are covered with a couple of most popular categories, which increases the risk of the classifier overfitting and underperforming on the long-tail product types.

A science challenge in e-commerce is the cold-start problem when expanding to a new locale; to cater for this case it is important to develop query understanding models that will be capable of adapting to the new marketplces without long and costly data collection and retraining.

Proposed approach and contributions. To overcome the mentioned issues with sparse and unbalanced data in low-resource locales, we propose to leverage rich data in high-resource locales via *knowledge transfer* for Q2PT problem.

We experiment with a unified multilingual multi-locale Q2PT model, which shares model parameters and training data across all stores. This solution allows small loacles to benefit from the knowledge of the established stores, solving the problem of data sparsity and increasing sample efficiency. Moreover, it reduces the training time and complexity, as well as memory requirements for storing a model checkpoint per each locale individually.

A possible disadvantage of this approach is that the unified *locale-agnostic* model can transfer biases from the large stores to the predictions on the low-resource ones [1]. To alleviate this, we propose a *unified locale-aware* model variant, by conditioning the prediction on the locale-id. The experiments show that locale-aware model performs better than the agnostic one, which demonstrates that it can better preserve locale-specific traits.

Further, we conduct extensive qualitative analyses on the query level, to assess how much the locales' differences affect Q2PT predictions. We find that Q2PT task is not locale-invariant (i.e. the same query yields different product type distributions depending on the country) and build categorization of local differences for product type predictions.

To support our findings we conduct large-scale experiments with an experimental e-commerce data set sampled from real data, including 20 locales and 1414 product types. Our insights from comparing non-unified and unified Q2PT models will be useful for the industrial and academic practitioners, helping to design query understanding systems with locale particularities in mind.

2 Related Work

Search query classification. E-commerce query understanding signals are essential for search rankers and other downstream services, and thus have received significant attention from the research community [6, 8, 9, 12, 15, 17–19].

An important challenge in Q2PT is classifying long-tail queries, which have scarce training data. Zhu et al. [18] improve long-tail query classification by transferring knowledge from the frequent queries, which are similar to them. The effects of query variation are studied in [17]; the authors notice that slight variation to the query can flip the category prediction, and they train the classification model to distinguish between pairs of queries, which share most of the terms but have different intent. This training strategy helps to overcome data sparsity for long-tail queries.

Multiple studies propose solutions to tailor existing classification models for short and ambiguous search queries. For instance, Liu et al. [9] propose a hybrid system which uses different models to serve long and short queries. Jiang et al. [6] develop a new pretraining task to improve over the standard term masking, which is detrimental for short texts: instead, they concatenate the input query with generated words and pretrain the model on identifying the extra terms.

Another issue is the shortage of query classification training data. To tackle this issue, Skinner and Kallumadi [15] propose to use transfer learning to infer query category using the model trained on product-category pairs. Alternatively Qiu et al. [12] propose generating synthetic queries, obtained from the product description substring, to pretrain the transformer models.

Finally, there are various other strategies proposed to improve search query classification. Zhu et al. [19] incorporate category contrastive loss into the model training, showing performance improvement for query classification with hierarchical classes. Zhang et al. [16] augment query category prediction by jointly training it with query-item semantic matching in a multi-task framework.

Handling data from multiple locales. Cross-market recommendations and search is an important problem for international e-commerce stores: the developed solutions need to meet the performance bar globally and be able to generalize to new locales. However, there has been only a few studies addressing this setup [1, 5, 13].

Bonab et al. [1] investigate market adaptation through fine-tuning a general model with limited data from a new locale. In this work, the product recommendation model is pretrained on all locales, but afterwards is fine-tuned on each store separately. Such approach helps the models to share some knowledge, but it complicates model training and serving, and incurs high storage costs. The method proposed in our paper, to fine-tune a unified model jointly on all locales, helps to overcome this problem.

In the area of multi-locale query classification, Lin et al. [7] develop a shard selection method for e-commerce product search. They conduct experiments under high-resource conditions, thus not addressing the low-resource issue in the smaller stores. Close to our work is [10], proposing a BERT-based query classification model, which has a separate product type classification layer for each locale. The unified backbone allows to reduce storage requirements, however, training the classification layer for each locale

separately reduces the model's ability to transfer knowledge. We use the architecture from Luo et al. [10] as one of the baselines for our experiments.

Similar to this study, Roitero et al. [13] compare the performance of global and locale-specific models in cross-market music recommendation domain, showing that market knowledge plays an important role.

3 Methodology

We study the task of predicting product type intent for a given e-commerce search query (Q2PT) in a multi-locale setting. This is a multi-label classification task: for an input query in each locale, the Q2PT model needs to return all product types associated with the query. In the multi-locale setup, each store has an independent catalogue of items, potentially a different language and different volumes of traffic.

In our study we propose an alternative to an existing method of training Q2PT models for each locale separately [10], replacing it with a unified architecture, which shares model parameters and training data across all stores.

We use BERT-based encoder [4] to create the representations of the input queries, following current research in query classification [10, 12, 18, 19]. The encoder is followed by a fully-connected layer with sigmoid activation, applied to the [CLS] token.

Building on this base architecture, we compare three variants:

- **non-unified (*NU*)** [10] - which has a common DistilBert backbone but separate classifiers for all locales. Each locale-specific classifier is trained only on the data from this specific store.
- **unified locale-agnostic (U_{ag})** - in this model both DistilBert encoder and classifier are shared across all locales and trained on the mixture of the global data.
- **unified locale-aware (U_{aw})** - this model is similar tho the locale-agnostic version, however, it conditions the product type prediction on the locale-id. To achieve it, we prepend the locale-id token to the input keywords, separated with a special [SEP] token.

Both unified model variants have their merits and drawbacks. U_{ag} solves the problem of data shortage in low-resource stores, but does not account for local specifics in different locales. Specifically, it can lead to biases in smaller locales, transferred from the bigger ones. U_{aw} overcomes this problem by encoding the input locale information alongside with the query. This effectively allows the model to produce different product type distributions for each locale.

On the other hand, locale-agnostic model has its advantages over locale-aware one: U_{ag} model is more practical for the cases of cold-start launches of new stores, with no prior training data. In this scenario U_{aw} needs to be retained to learn about new locale-id, while U_{ag} can be readily used out of the box.

4 Experimental Setup

4.1 Datasets

We use an experimental data set sampled from real data, covering 1414 product types and 20 locales: (**US** - United States, **DE** - Germany, **UK** - United Kingdom, **JP** - Japan, **IN** - India, **IT** - Italy, **CA** - Canada, **FR** - France, **ES** - Spain, **MX** - Mexico, **BR** - Brazil, **AE** - United Arab Emirates, **AU** - Australia, **SA** - Saudi Arabia, **EG** - Egypt, **NL** - Netherlands, **TR** - Turkey, **SE** - Sweden, **SG** - Singapore, **PL** - Poland). To the best of our knowledge, there are no existing studies benchmarking query classification models across this large number of locales.

4.2 Training data

The data for training Q2PT models is created by aggregating fully anonymized customer click-through behavior, following previous research [7, 9, 17]. The product type for the user search query is derived as the majority product type of items, that the user clicked following that query. Formally, we define the probability of the query q belonging to the product type p as follows:

$$P(q, p) = \frac{num\_clicks(q, item | item \in p)}{num\_clicks(q, item)} \quad (1)$$

After that we select all <*query, product type*> pairs that have $P(q, p) > 0.5$.

The collected data has immense discrepancy in sample distribution among the locales: the large stores take over 90% share of the dataset. We split the locales into 2 buckets: *High-Resource* (Hi-Re) locales, including US, DE, UK, JP, IN, IT, CA, FR, ES stores, and *Low-Resource* (Lo-Re) locales, including MX, BR, AE, AU, SA, EG, NL, TR, SE, SG, PL stores, based on the number of training samples. In our experiments we compare the models' results per bucket, to assess the effects of unified and disjoint training for these two groups.

We use 100s of millions samples for model training and 10s of millions for validation and hyperparameter selection.

4.3 Evaluation Data

As the customer click-through data is noisy and prone to trends and seasonality, we cannot rely on it for accurate model performance evaluation. Instead, we create two separate evaluation datasets: human-annotated and automatically weakly labeled.

Human-labeled data. We recruited professional annotators to label search queries with all applicable product types; for each locale we got around 1k human-labeled queries. The human-annotated dataset is high-quality, however, it mostly consists of queries associated with popular PTs. Thus, this dataset has a very low coverage of the product type label space: on average around 600 product types (42% of the whole list) are included for each locale, moreover, on average only 22 PTs are associated with at least 20 labeled samples.

As an alternative, to be able to conduct a more fine-grained per-PT anaysis, we created a large-scale automatically labeled dataset.

Automatically labeled data. To create this dataset we leverage relevance labels for the <*query, item*> pairs, obtained from a pre-trained classifier. For each query we collect the categories of all items that are predicted relevant to it. The query label is then selected as a majority category from relevant items.

The resulting dataset consists of 2.7M labels in total, with an average of 120k labels per locale. To check the correctness of this labelling, we manually inspected 200 queries from different locales, and verified that this labeling method achieves almost 90% accuracy. Importantly, this method allowed us to collect a sufficient number of evaluation samples for all 1414 product types.

	MX	BR	AE	AU	SA	EG	NL	TR	SE	SG	PL	Lo-Re
NU	0.91	0.81	0.9	0.84	0.83	0.72	0.66	0.87	0.53	0.91	0.54	0.77
$\mathbf{U}_{ag}$	0.92	0.78	0.9	0.85	0.87	0.77	0.69	0.9	0.55	0.91	0.59	0.79 (+2%)
$\mathbf{U}_{aw}$	0.92	0.82	0.91	0.87	0.87	0.76	0.70	0.90	0.58	0.92	0.60	0.80 (+3%)

	US	DE	UK	JP	IN	IT	CA	FR	ES	Hi-Re	WW
NU	0.79	0.9	0.88	0.76	0.85	0.93	0.92	0.89	0.92	0.87	0.82
$\mathbf{U}_{ag}$	0.81	0.91	0.9	0.79	0.83	0.95	0.93	0.91	0.93	0.89 (+2%)	0.83 (+1%)
$\mathbf{U}_{aw}$	0.80	0.91	0.90	0.80	0.86	0.95	0.92	0.91	0.93	0.89 (+2%)	0.84 (+2%)

Table 1: Results for recall at 0.8 precision, comparing the baseline (*NU*) and the unified methods ($\mathbf{U}_{ag}$, $\mathbf{U}_{aw}$), in Low-Resource (top) and High-Resource (bottom) locales on the human-labeled evaluation set.

Nevertheless, the automated approach has the following drawbacks: i) by taking the majority product type label, we make an assumption that the query can be associated with only one product type. This is a valid assumption in most cases, but there are some exceptions (e.g. a query '*sun protection*' can mean a *cream*, *apparel* or *glasses*); ii) the quality of the resulting annotations depend on the quality of the relevance classifier; iii) the distribution of the queries and product types in this dataset does not match the one in the training data.

Taking into account these drawbacks of the automated approach, we opt to use human-annotated queries as our main evaluation data, and use the automatically-labeled one for an additional in-depth per-PT analysis.

4.4 Model Implementation and Training

We used a multilingual pretrained DistilBert [14] checkpoint, which was fine-tuned on e-commerce queries; we further fine-tuned the model on the Q2PT task. We used Adam optimizer and trained the model with 8e-5 learning rate, 0.001 dropout and 2^{11} batch size until convergence, using binary cross-entropy loss. The hyperparameters were chosen through grid search on the validation split.

4.5 Evaluation Metrics

We report recall at 0.8 precision: the fixed high-precision setting is a standard in evaluating customer-facing applications, such as e-commerce sites. Given the importance of Q2PT signal for the downstream search and recommendation components, the query classification models have to meet high precision bar.

5 Experiment Results

Results on the human-labeled data. In Table 1 we report recall at 0.8 precision for all 20 locales. Additionally, we aggregate the results separately for established High-Resource (Hi-Re) and emerging Low-Resource (Lo-Re) stores, and worldwide (WW).

We observe that both variants of the unified model outperform the non-unified one for all locales, with the total increase of 2% recall worldwide. Notably, the most benefiting locales are small ones, e.g. PL (+6%) and SE (+5%). It shows that the unified models can efficiently transfer knowledge from high-resource to the low-resource locales. Importantly, we notice that even on the Hi-Re locales a generalist U_{aw} model performs slightly better than a specialist U_{aw} model.

Between the two variants of a unified model, U_{aw} has slightly more pronounced gains over U_{ag}. This illustrates that the task of product type prediction is not locale-invariant: conditioning on the locale information helps the model to distinguish locale-specific peculiarities (we discuss it further in Section 6.1).

To further assess the models on low- and high-resource locales we plot precision-recall curves for one of the high-resource (US) and one of the small-resource (PL) locales in Figure 2. For US, U_{aw} slightly dominates for the high precision values, thanks to the both more diverse training data from the other locales and preserved information about the current locale. In PL, there is a significant gap between *NU* and consolidated models, because the latter were trained on 1000 times more data.

Additionally, we experimented with a completely disjoint model architecture: for each locale all model parts are shared among the locales, and they are trained and stored separately. We found that this variation performs on par with U_{aw} and has 2% performance increase compared no *NU*. Given that this model takes 7 times longer to train (assuming sequential training on all locales) and 20 times greater memory requirements, without significant performance improvements, we did not consider this model for our analyses.

Results on the automatically-labeled data. We compute recall at 0.8 precision on the automatically-labeled evaluation set, and present the aggregated results in Table 3. On this dataset the gap between *NU* and unified models considerably increases, which can be attributed to the automated data having more long-tail product types, which are challenging to predict, and which have extremely scarce training data in low-resource locales. At the same time, human data is largely composed of the most popular product types, on which the performance of all models is on par.

In this dataset, however, the trend between both unified models changes, with U_{ag} having slightly better performance. After examining the performance of the models per locale, we found that U_{ag} has superior results in all but two of the biggest locales: US and UK (in these locales U_{ag} performance is 1% worse than U_{aw}). It shows that conditioning on locale-id helps the model to develop some locale-specific knowledge, which is abundant in high-resource locales, and thus improves the performance of U_{aw}.

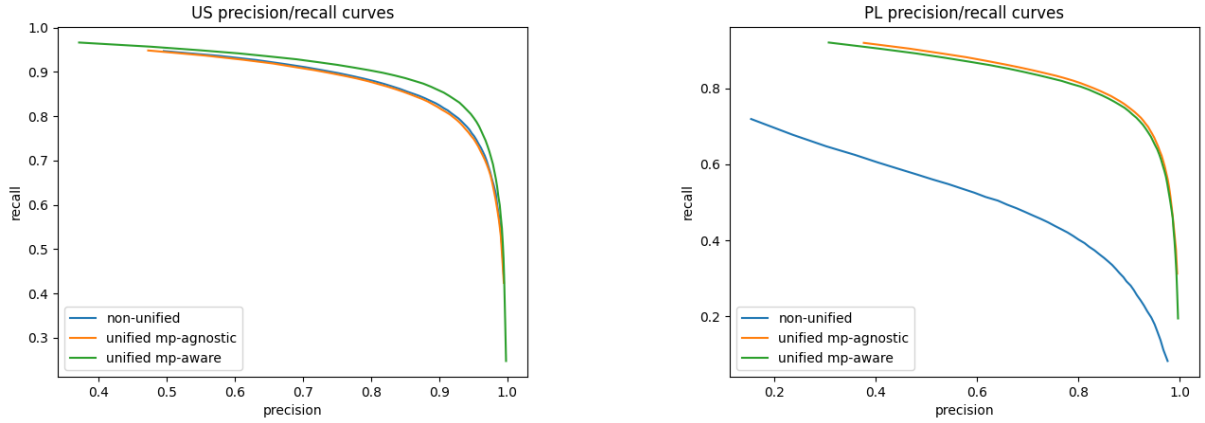


Figure 2: Precision-recall curves in US (left) and PL (right) locales.

	Lo-Re	Hi-Re	WW
NU	0.64	0.80	0.71
$\mathbf{U}_{ag}$	0.83 (+17%)	0.87 (+7%)	0.85 (+14%)
$\mathbf{U}_{aw}$	0.81 (+15%)	0.86 (+6%)	0.83 (+12%)

Table 2: Results for recall at 0.8 precision, comparing the baseline (*NU*) and the unified methods ($\mathbf{U}_{ag}$, $\mathbf{U}_{aw}$), on the automatically labeled evaluation set.

	WW		Lo-Re		Hi-Re	
	NU	$\mathbf{U}_{aw}$	*NU*	$\mathbf{U}_{aw}$	*NU*	$\mathbf{U}_{aw}$
correlation	0.2	0.15	0.22	0.15	0.16	0.14
head	0.85	0.9	0.83	0.89	0.87	0.91
torso	0.79	0.85	0.77	0.84	0.82	0.87
tail	0.7	0.8	0.65	0.78	0.75	0.83

Table 3: Pearson correlation of the number of samples and accuracy per-PT (top line), accuracy per product type on the head/torso/tail PT splits.

Qualitative analysis. We also note that the knowledge transfer can have negative effect on the Q2PT predictions in some cases. To validate it, we checked the evaluation queries for which the *NU* model made a correct prediction, while a unified model (we use U_{aw} in this study) made an error. For instance, the query '*roma*' should yield different product types depending on the store: it should return *laundry_detergent* for MX (local brand), *personal_fragrance* for DE and other European stores (name of perfume popular in Europe), and for the remaining locales the correct prediction will be a *book*. We notice that in SG U_{aw} model, biased by the data from the other locales, predicts *laundry_detergent* for the query '*roma*', while *NU* correctly predicts *book*.

Another interesting dimension of analysis is the differences between the errors of U_{ag} and U_{aw} on the automatically-labeled dataset. Given that US and UK are the only locales where U_{aw} outperforms the locale-agnostic model, we investigated the US queries where U_{ag} fails, while U_{aw} is correct. One example query is '*boys drawers*': while 'drawers' is a common name for underwear in Asia, in the US this term is not used, and the locale-aware model predicts the correct product type as *dresser*. At the same time, U_{ag} model, which does not have access to the locale information, erroneously predicts PT *underpants*.

Another example is query '*vanity side shelf with drawers*': in the US the term 'vanity', among other things, refers to furniture, while in British English this is uncommon. Therefore, the products categorized as *makeup_vanity* in US, have *table* product type in other stores. This influences the training data for the models, which is based on the user clicks on the product categories, as described in Section 4.2. Therefore, in this case for the query '*vanity side shelf with drawers*' in US locale U_{aw} model predicts the product type *makeup_vanity*, while U_{ag} predicts *table*.

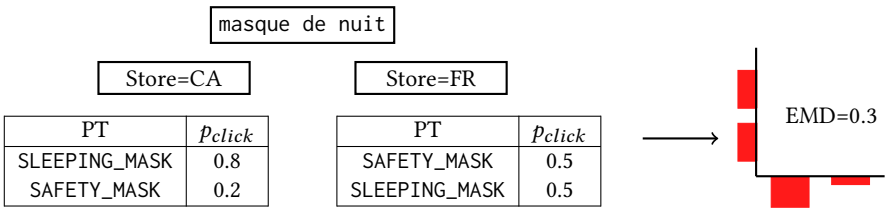
Finally, we note that such cases of meaningful discrepancies are rare (less than 1% of the automatic evaluation set), as mostly the differences in the models' predictions arise from one of the models refraining from prediction.

5.1 Analysis on the Product Type Level

In this analysis we aim to evaluate the models on different product type groups, based on their frequency in the training data. As shown in Figure 1, the distribution of product types is very skewed, and thus it is desirable to assess the models' capabilities to correctly infer long-tail product types.

For this analysis we split the list of PTs into 3 parts: *head* (very frequent PTs, e.g. *book*), *torso* (average frequency, e.g. *wall ornament*) and *tail* (niche PTs, e.g. *mounted storage system kit*), so that each part has 1/3 of query mass of the whole dataset. As a result we got 48 head, 198 torso and 1168 tail product types.

For this experiment we use automatically-labeled data, because it evenly covers all product types. We compare the performance of non-unified model against unified locale-aware model, as it has shown superior performance to the U_{ag} on the human-labeled benchmark. We compute per-PT accuracy, which is defined as the number of correctly predicted occurrences of a product type, over all occurrences of this product type in the evaluation set. Additionally, we compute Pearson correlation of the PT frequency and PT



PT	p_{click}
SLEEPING_MASK	0.8
SAFETY_MASK	0.2

PT	p_{click}
SAFETY_MASK	0.5
SLEEPING_MASK	0.5

Figure 3: The query "*masque de nuit*" exists in both training sets for CA and FR. In CA, the most clicked PT is `SLEEPING_MASK` while in FR, `SAFETY_MASK` and `SLEEPING_MASK` have equal rates, which results in EMD of 0.3

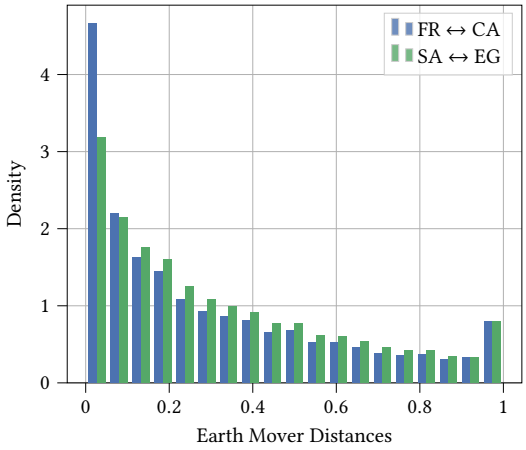


Figure 4: Density of EMD distances for queries in the intersection of FR ↔ CA and SA ↔ EG pairs.

accuracy, to see how much the models overfit on the high-frequency product types.

The results of this analysis are shown in Table 3, aggregated for low- and high-resource locales and worldwide.

Unsurprisingly, both models perform better on the more frequent groups of product types, however, the accuracy drop from head to tail PT group is more prominent for the non-unified model. Additionally, we noted that the unified model dramatically reduces the number of product types, which don't meet a high accuracy bar of at least 0.5. Non-unified model has 105 such low-accuracy PTs, compared to 27 for a unified model. Most of those 27 PTs are related to digital content, which are easy to confuse to each other (e.g., *music track* and *music album*).

Consistently with the previous results, the correlation of accuracy and PT frequency is greater in non-unified model. One interesting observation is that the correlation difference is small on Hi-Re locales and is more pronounced on the Lo-Re stores. Additionally, we compared correlations in the biggest locale (US) and the smallest (PL). The correlation in US was similarly low for both NU and U_{aw} (0.12 and 0.9 resp.), however, in PL the correlation of NU accuracy is extremely high (0.23) and it is getting smoothed with U_{aw} (0.11).

6 Discussion

6.1 Analysis of Locale Differences

In our experiments we observe that the same search query might have different meanings depending on the locale, and thus yield different product type distributions. We conduct additional analyses to quantify how frequent this phenomenon occurs.

We focus on two pairs of locales that share the same language: FR ↔ CA (French) and EG ↔ SA (Arabic); for each pair of stores, we consider the queries that exist in the training sets in both markets and measure the difference between per-PT click distributions (from Equation 1). To this end, we compute the classic Earth Moving Distance (EMD) [2, 11]. In Figure 3 we illustrate one example of such computation.

We depict the histograms of EMD distributions for each locale pair in Figure 4: the smaller the EMD, the more similar the locale per-PT distributions are. We observe an expected peak near zero, because largely the customers in different stores should have the same PT in mind in their search. However, the amount of dissimilar modes is significant, ranging from small differences to completely different distributions (EMD near 1).

6.2 Categorization of Local Differences

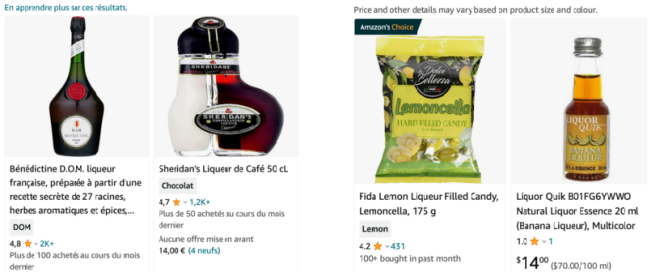
During analysis of query differences, we found multiple cases where the product type distributions significantly vary across locales. We attribute those discrepancies to one of the following cases:

1. Dialectal differences: despite the same language, the query has a different meaning to the users in different stores. One particular example is the word '*liqueur*', which means an alcoholic drink in France, but a non-alcoholic drink or syrup in Canada (see Figure 5a). Another example is '*vaporizer*', which means a smoking gadget in France and an air humidifier appliance in Canada (depicted in Figure 5b). These discrepancies can be explained by cultural and historical factors, e.g. language drift from the neighbouring countries. Although such cases are pretty rare in the data, they can have a profound impact on the user experience, especially in the cases of products under legal regulations.

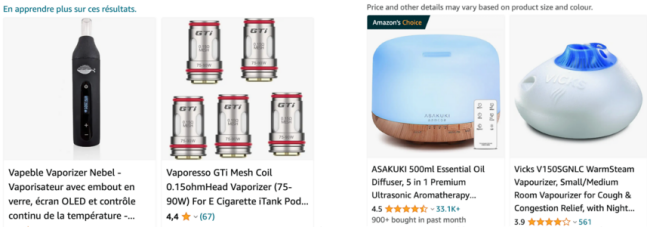
2. Selection differences: the query has different meaning with respect to the product type, that is caused by a mismatch in the selection of products offered in corresponding markets. For example, a query "carpe" that means "a carp" in French leads to fishing accessories in FR store. However, in Canada it is mapped to a popular cosmetic brand marketed under the same name (see Figure 5c).

3. Noisy differences: in many cases we observe a large discrepancy for a given query as measured by EMD between PT click

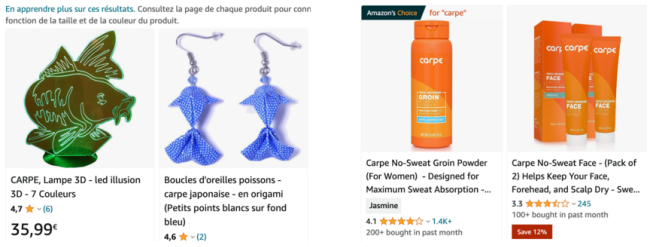
distributions. However, this metric is not always conclusive since EMD doesn't account for uncertainties in click probabilities from Equation 1. As a toy example, a PT with 5 clicks and 10 impressions should be treated differently than PT with 10K clicks and 20K impressions, even though the probability p will be equal to 0.5 in both cases.



(a) `liquere` products



(b) `vaporizer` products



(c) `carpe` products

Figure 5: Examples of product type discrepancies between FR (left 2 products) and CA (right 2 products) stores.

7 Conclusion and Future Work

In this study we investigate the task of e-commerce query product type prediction in a multi-locale setup and propose a transfer learning solution to augment Q2PT predictions in low-resource stores to achieve global parity.

The evidence from our offline experiments and user studies shows that the unified Q2PT model, sharing the parameters and training data across all locales, outperforms the non-unified model, on both low- and high-resource locales, additionally decreasing infrastructure requirements. We compare locale-agnostic and locale-aware variants of the unified model, showing that it is important to capture store-specific characteristics by conditioning the prediction on the locale-id. The findings from our work will be useful for practitioners developing multi-locale query classification models.

We identify the following avenues for our future work. We plan to investigate the impact of language variation on the Q2PT task, both in terms of the language of the input query as well as the model's pre-training data. Additionally, we plan to conduct similar analyses for other related query understanding models, such as brand classifiers.

References

[1] Hamed Bonab, Mohammad Aliannejadi, Ali Vardasbi, Evangelos Kanoulas, and James Allan. 2021. Cross-market product recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 110–119.

[2] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. 2011. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*. 1–12.

[3] Yi Chang and Hongbo Deng. 2020. *Query understanding for search engines*. Springer.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Bruce Ferwerda, Andreu Vall, Marko Tkalcic, and Markus Schedl. 2016. Exploring music diversity needs across countries. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 287–288.

[6] Haoming Jiang, Tianyu Cao, Zheng Li, Chen Luo, Xianfeng Tang, Qingyu Yin, Danqing Zhang, Rahul Goutam, and Bing Yin. 2022. Short text pre-training with extended token classification for e-commerce query understanding. *arXiv preprint arXiv:2210.03915* (2022).

[7] Heran Lin, Pengcheng Xiong, Danni Danqing Zhang, Fan Yang, Ryoichi Kato, Mukul Kumar, William Headden, and Bing Yin. 2020. Light feed-forward networks for shard selection in large-scale product search. (2020).

[8] Yiu-Chang Lin, Ankur Datta, and Giuseppe Di Fabbrizio. 2018. E-commerce product query classification using implicit user's feedback from clicks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 1955–1959.

[9] Xianjing Liu, Hejia Zhang, Mingkuan Liu, and Alan Lu. 2019. System Design of Extreme Multi-label Query Classification using a Hybrid Model.. In *eCOM@ SIGIR*.

[10] Chen Luo, William Headden, Neela Avudaiappan, Haoming Jiang, Tianyu Cao, Qingyu Yin, Yifan Gao, Zheng Li, Rahul Goutam, Haiyang Zhang, et al. 2022. Query attribute recommendation at amazon search. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 506–508.

[11] Gaspard Monge. 1781. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.* (1781), 666–704.

[12] Yiming Qiu, Chenyu Zhao, Han Zhang, Jingwei Zhuo, Tianhao Li, Xiaowei Zhang, Songlin Wang, Sulong Xu, Bo Long, and Wen-Yun Yang. 2022. Pre-training tasks for user intent detection and embedding retrieval in e-commerce search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4424–4428.

[13] Kevin Roitero, Ben Carterette, Rishabh Mehrotra, and Mounia Lalmas. 2020. Leveraging behavioral heterogeneity across markets for cross-market training of recommender systems. In *Companion Proceedings of the Web Conference 2020*. 694–702.

[14] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[15] Michael Skinner and Surya Kallumadi. 2019. E-commerce Query Classification Using Product Taxonomy Mapping: A Transfer Learning Approach.. In *eCOM at SIGIR*.

[16] Hongchun Zhang, Tianyi Wang, Xiaonan Meng, Yi Hu, and Hao Wang. 2019. Improving Semantic Matching via Multi-Task Learning in E-Commerce. *eCOM@ SIGIR* 10 (2019).

[17] Junhao Zhang, Weidi Xu, Jianhui Ji, Xi Chen, Hongbo Deng, and Keping Yang. 2021. Modeling Across-Context Attention For Long-Tail Query Classification in E-commerce. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 58–66.

[18] Lvxing Zhu, Hao Chen, Chao Wei, and Weiru Zhang. 2022. Enhanced representation with contrastive loss for long-tail query classification in e-commerce. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. 141–150.

[19] Lvxing Zhu, Kexin Zhang, Hao Chen, Chao Wei, Weiru Zhang, Haihong Tang, and Xiu Li. 2023. HCL4QC: Incorporating Hierarchical Category Structures Into Contrastive Learning for E-commerce Query Classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3647–3656.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009