Cross-lingual Search for e-Commerce based on Query Translatability and Mixed-Domain Fine-Tuning

Jesus Perez-Martin*
jesus.perez-martin@walmart.com
Walmart Global Tech
Sunnyvale, CA, USA

Jorge Gomez-Robles*
jorge.gomez.robles@walmart.com
Walmart Global Tech
Sunnyvale, CA, USA

Asier Gutiérrez-Fandiño
asier.gutierrez-fandino@walmart.com
Walmart Global Tech
Sunnyvale, CA, USA

Pankaj Adsul
pankaj.adsul@walmart.com
Walmart Global Tech
Sunnyvale, CA, USA

Sravanthi Rajanala
sravanthi.rajanala@walmart.com
Walmart Global Tech
Sunnyvale, CA, USA

Leonardo Lezcano
leonardo.lezcano@walmart.com
Walmart Global Tech
Sunnyvale, CA, USA

ABSTRACT

Online stores in the US offer a unique scenario for Cross-Lingual Information Retrieval (CLIR) due to the mix of Spanish and English in user queries. Machine Translation (MT) provides an opportunity to lift relevance by translating the Spanish queries to English before delivering them to the search engine. However, polysemy-derived problems, high latency and context scarcity in product search, make generic MT an impractical solution. The wide diversity of products in marketplaces injects non-translatable entities, loanwords, ambiguous morphemes, cross-language ambiguity and a variety of Spanish dialects in the communication between buyers and sellers, posing a thread to the accuracy of MT. In this work, we leverage domain adaptation on a simplified architecture of Neural Machine Translation (NMT) to make both latency and accuracy suitable for e-commerce search. Our NMT model is fine-tuned on a mixed-domain corpus based on engagement data expanded with catalog back-translation techniques. Beyond accuracy, and given that translation is not the goal but the means to relevant results, the problem of *Query Translatability* is addressed by a classifier on whether the translation should be automatic or explicitly requested. We assembled these models into a query translation system that we tested and launched at *Walmart.com*, with a statistically significant lift in Spanish GMV and an nDCG gain for Spanish queries of +70%.

CCS CONCEPTS

• **Information systems** → **Multilingual and cross-lingual retrieval**; Online shopping; • **Computing methodologies** → *Machine translation*; *Supervised learning by classification.*

KEYWORDS

neural machine translation, product search, cross-lingual information retrieval, domain adaptation, query classification, fine-tuning, back-translation

ACM Reference Format:
Jesus Perez-Martin, Jorge Gomez-Robles, Asier Gutiérrez-Fandiño, Pankaj Adsul, Sravanthi Rajanala, and Leonardo Lezcano. 2023. Cross-lingual Search for e-Commerce based on Query Translatability and Mixed-Domain Fine-Tuning. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3543873.3587660

1 INTRODUCTION

There is a growing demand for B2C e-commerce search engines to address language barriers and cultural differences [32]. To improve query understanding, precision and recall [1, 25, 26, 36], recent approaches [12, 34] have explored the automatic translation of user queries as one of the first steps in the search. This approach is specially important for platforms that serve a global audience, as these systems must be able to handle a wide range of languages and cultural differences. That is the case of online stores and marketplaces in the US, where Hispanics represent approximately 18% of total population, *i.e.*, more than 60 millions [9].

While physical stores allow non-English speaking customers to visually find a product, the e-commerce search experience requires familiarity with specific English terms. With about 30% of Hispanics speaking Spanish at home, such terminology is frequently lacking. Moreover, 40% of consumers refuse buying from websites in other languages [32]. We note that online product search from Spanish to English in the US poses the following problems that render generic MT insufficient:

(**P1**) **Latency**: For search engines serving millions of customers, online translation speed is as important as accuracy. A single slip-up could make the difference in the competitive e-commerce space.

(**P2**) **Code-mixing**: The mix of languages in a single query poses a challenge for search. Colloquially known as *Spanglish*, Hispanic customers in the US frequently mix Spanish and English in the same query; *e.g.*, "*cake de fresa*", "*display grande*". If the Spanish content is not detected and translated, the search precision and recall are limited to the understanding of the English part.

(**P3**) **Ambiguity**: The lack of context also harms language detection; *e.g.*, while in English the "*pan*" relates to *cookware*, it means

*Both authors contributed equally to this research.

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587660

bread in Spanish. This is cross-language ambiguity. In addition, there are Spanish terms like "*calabaza*" that can be translated to more than one in English (*pumpkin* and *butternut squash*).

(P4) Query Translatability: Offering the most relevant result requires not only detecting Spanish but also deciding whether to automatically translate or not. For example, if translated, the Spanish query "*recetas de cocina*" (*cooking recipes*) would lose its implicit Spanish intention, and "*gran turismo*" (*grand tourer*) would not retrieve the expected video game.

(P5) Non-Translatable entities: As opposed to generic text, product-oriented queries carry a high density of brands, named entities, loanwords, sport teams, etc. For example, "*corona*" (*crown*) should not be translated when referring to the *beer's* brand and "*queso blanco*" (*white cheese*) would not imply the Latin American style of cheese.

(P6) Spanish dialects: With Hispanic migration from over 20 Spanish-speaking countries to the US, the same intention is sometimes expressed through heterogeneous terms. For example, there are 15+ Spanish words for *popcorn*, and 6+ for *appetizer*.

The presented CLIR experience tackles the above listed problems in an e-commerce setting where both accuracy and efficiency of translations are required. Spanglish queries are detected and translated to English, enabling the retrieval of the most relevant results from a monolingual search engine.

We also propose a complete methodology for creating a mixed-domain corpora to fine-tune small NMT models. This methodology ensures a diverse domain-specific language, allows to have better control over the characteristics of the corpus, and ensures its quality and consistency. We assess the effectiveness of our solution upon *Walmart.com* search.

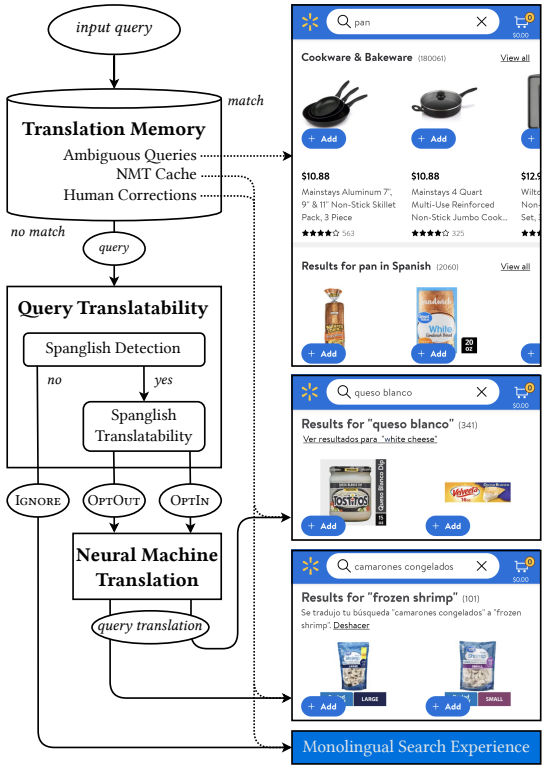

Figure 1: Query translation experience at *Walmart.com*. The screenshots on the right show the experiences we propose.

2 RELATED WORK

Previous CLIR efforts in e-commerce have used MT to bring the user query into the search system's primary language [12, 34]. To deal with the latency requirement (see **P1**), Yao et al. [34] proposed an asynchronous strategy combining the online speed of Statistical Machine Translation (SMT) with the offline correctness of NMT. In recent years, frameworks specifically designed for fast MT have been developed [15, 20, 21]. In Section 5, we propose to use *Marian-NMT* [15] which is an efficient and highly-optimized NMT framework [16, 23] written in C++ with minimal dependencies.[1] The use of Marian-NMT has allowed us to efficiently fine-tune our NMT model on a large amount of data and deploy it in production for synchronous translations at run-time.

The problem of query translatability has been addressed elsewhere only from the perspective of language detection [14, 22]. In addition, we emphasize the need to retain the original language when the query carries a language specific intention. For example, "*recetas de cocina*" not only is Spanish but also implicitly implies that Spanish books and content are expected in the results. If translated to "*cooking recipes*", there is a drastic drop in precision. An automatic vs on-demand translation classifier to solve this problem is documented in Section 4.2.

Hu et al. [12] proposed a synthetic nDCG approach that relies on the items previously purchased through the non-translated query to asses the relevance of translation. We find that the reasons for which translation is needed in the first place also impair the accuracy of this nDCG approach: (i) for non-translated queries that suffer from low recall, the synthetic nDCG can only report partial lifts; (ii) for the ones that lead to irrelevant inventory, the synthetic nDCG will not reach significance; and (iii) for the non-translated queries that are biased towards item titles enriched with Spanish tokens, the synthetic nDCG would also be biased. To avoid such limitations, the CLIR experience we present in this paper was assessed through a manual nDCG evaluation. It was carried by bilingual judges that rated the query-item pairs independently from previous purchases.

3 CROSS-LINGUAL IR OVERVIEW

We present an efficient query translation system that receives the user query and returns a *translatability class* to drive the CLIR experience and a *translated query* to retrieve results from the underlying search engine. The right side of Figure 1 shows the four experiences in ascending order of popularity: **Ambiguous** are rare queries where the user is offered multiple carousels, as relevant results may come from both translating and not translating, or from different translations; **OptIn** for queries that are Spanish but require an explicit signal from the user to be translated; **OptOut** for most of Spanish queries, which benefit from automatic translation, offering users the option to withdraw; and **Ignore**, which vastly dominates traffic as it includes all the English queries that are not to be translated.

[1] Marian-NMT website: marian-nmt.github.io

As shown in Figure 1, the Translation Memory is the first back-end module. It caches the resulting experience and translations for every Spanish query reaching the system so that repeated calls get a very low lookup latency. Given the skewness in search behavior, the cache can serve approximately 80% of Spanish traffic, which is comparable with the 90% reported by Yao et al. [34]. If there is a memory miss, the Query Translatability module described in Section 4 decides the user experience and the NMT model presented in Section 5 generates the translation. The memory also stores ambiguous queries (see **P3**). Ambiguity raises in search when query polysemy can not be solved by context. Two types need to be addressed for proper Spanish query understanding:

Cross-Language Ambiguity occurs when there are drastically different purchases for a Spanish query, between OptIn and OptOut. The difference is measured as the distance across hierarchical departments. To support both intentions, two UX carousels of horizontal items are presented to the user as shown for "*pan*" in Figure 1. **Spanish Ambiguity** occurs when two different and popular English queries translate to the same Spanish query. To disambiguate, the two English queries are used to retrieve items in the two separated UX carousels.

As they are rare exceptions, the ambiguous queries can be detected offline and stored in the Translation Memory depicted in Figure 1. They can go back to single intention if one of the carousels drives significantly more engagement than the other.

4 QUERY TRANSLATABILITY

As the first sub-module of Translatability, the role of Language Detection (Section 4.1) is to Ignore the large percentage of English traffic, allowing for the Spanish Translatability model (Section 4.2) and NMT model (Section 5) to focus on Spanish queries only.

4.1 Spanglish Language Detection

We detect as Spanglish (see **P2**) the queries where at least one token is in a lexicon of 280K Spanish terms sampled from *wiktionary.org*, which includes regional variations, dialects (see **P6**), and unlike a standard dictionary, it is constantly updated by volunteers.

This approach yields better results than the alternative English lexicon matching and the pre-trained language classifier proposed by Joulin et al. [14]. On the one hand, the English lexicon strategy is prone to false negatives since tokens in a query could all match the English lexicon but intend an equally spelled Spanish word depending on the context in which they are used. For instance, in the queries "*pan light*" and "*bates for kids*," the words *pan* and *bates* are more likely to be the Spanish for *bread* and *baseball bats* and not the English words referring to *cookware* and *furniture*. One the other hand, we found that general-purpose language models may output low Spanish probabilities for search queries in Spanglish. For example, by using the model proposed by Joulin et al. [14] we would fail to detect the queries "*frijoles*" and "*display grande*". Compared to it, our method improves the detection of Spanish words by 23% and Spanglish queries by 55%.

4.2 Spanglish Translatability

Whether a query has to automatically be translated or not is key to satisfy the query intent (see **P4**). For example, the query "*gran turismo*" is the title of a famous video game and thus, it should not be translated to "*grand tourer*" unless explicitly indicated by the user. Similar behavior is expected for queries related to books and music albums in Spanish. We address this problem with a model that learns to classify such queries as OptIn and the remaining ones as OptOut. We explain our approach and results below.

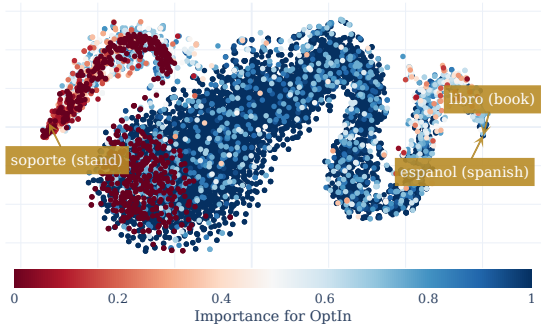

Figure 2: t-SNE projections of 10K word embeddings sampled at random. Given a set Q_t of queries associated to a token t, the importance score is computed as $\frac{|\{q \in Q_t | \textbf{label}(q) = \textsc{OptIn}\}|}{|Q_t|}$.

4.2.1 Translatability Dataset. We collect queries in Spanglish from our database and annotate as OptIn those leading to the selling of books, songs, albums or video games with titles in Spanglish too. For every token in the OptIn samples, we search for the most popular counter-examples; *i.e.*, queries with one common token but that do not lead to the selling of items in any of the OptIn categories. We guarantee a high number of counter-examples by sampling twice the queries for any token. For example, if *turismo* appears only once in the OptIn query "*gran turismo*", we search for two counter-examples having *gran* but not *turismo*, and two having *turismo* but not *gran*. The resulting dataset is of size n = 308K.

4.2.2 Model training. We train a binary *fastText* classifier [14] from scratch on a stratified split containing 66.7% of the samples. To overcome the inherent imbalance of the data, we augment this set by duplicating the instances and their variations without stop words nor numbers. To obtain the best training configuration, we apply Bayesian Optimization with SMAC3 [24] on the hyperparameters in Table 2. A total of 100 configurations are explored with the objective function being the 5-fold cross-validated accuracy on the augmented training set. Ultimately, a final model is yielded using the best hyperparamters on the whole training set.

4.2.3 Model evaluation. Table 1 shows the performance of the model on the remaining 33.3% of the data. Figure 2 shows the embedding space learned by our model, where tokens that are usually together in OptIn queries are separated from those only associated to OptOut queries. For example, the embeddings for *libro* (*book*) and *espanol* (*spanish*) are close to each other (right side) because they are together in many OptIn queries; whereas *soporte* (*stand*) is farther away (left side) so that a query such as *"soporte libro"* (*book stand*) gets classified as OptOut. This separation is learned because the *negative sampling* hyperparameters force the model to contrast OptIn queries to their counter-examples, which highlights

Table 1: Translatabily dataset and model performance. Data composition shows the proportion of each class in the split. Performance is reported on the test split.

Class	Dataset composition			Performance metrics		
	Full	Training	Test	Precision	Recall	F1
OptIn	14.36%	48.18%	14.78%	97.92	92.59	95.18
OptOut	85.64%	51.82%	85.22%	98.73	99.66	99.19
Macro average	-	-	-	98.32	96.13	97.19

Table 2: *fastText* hyperparameters and search space. Best configuration found in the 37th iteration.

Hyperparameter	Description	Search space	Best
dim	The dimension d of the word embeddings.	[50, 150]	91
wordNgrams	The number N for word N-grams.	[1, 5]	5
minn, maxn	The range of values for the char n-grams.	[1, 5]	2, 3
ws	The size of the context window.	[1, 5]	3
lrUpdateRate	The number of iterations between updates.	[100, n]	247381
lr	The learning rate.	[0.5, 1]	0.91
epochs	The number of training epochs.	[15, 25]	16
loss	The loss function.	{ns, softmax, hs}	ns
neg	The number negative samples.	[1, 10]	8

the importance of having this duality in the data. Complementary, the values of *wordNgrams* and *ws* enable the model to learn from large groups of words instead of individual tokens only. Particularly, if we set *wordNgrams* to 1 in the best configuration, the F1 score in the OptIn class drops to 85.95.

5 NEURAL MACHINE TRANSLATION

For NMT to be feasible for online translation, a lightweight network is needed (see **P1**). We adopt Tiny.Untied [3], an *encoder-decoder* architecture [5]. It uses the Transformer's [33] encoder with six attention-based blocks. The decoder is a stack of only two SSRU-untied layers [18]. Its vocabulary $\mathcal{V}$ is of dimension $|\mathcal{V}| = 32,000$, the embeddings of size 256, and filters of size 1536, resulting in a low cardinality of parameters, $|\theta| = 16.9$M.

We selected Tiny.Untied as the baseline for the domain adaption described in next sections after comparing its performance to larger and well established models: OpusMT [31], m2m100_418m [8] and BergamotLarge [3]. We report in Table 3 the metrics BLEU [27] and CHRF [28] for token and character matching, and BERTscore [35], which is based on token similarity using contextual embeddings. We compute these scores on our test set (Section 5.2).

For the NMT model adaption, we propose a mixed-domain-based fine-tuning for adapting the NMT model to our specific domain. During the fine-tuning process, the model is exposed to both general-domain and specific-domain data simultaneously. The general-domain data helps the model to retain its ability to translate more general phrases and sentences, while the specific-domain data helps it learn the unique terminology and code-mixed style [6, 10, 29] of the e-commerce domain.

5.1 Mixed-domain Corpora for Fine-Tuning

Given a set $\mathcal{D}$ of pairs $(q_{\text{SRC}}, q_{\text{REF}})$, where $q_{\text{SRC}} = (x_1, x_2, \ldots, x_n)$ is an input query, and $q_{\text{REF}} = (y_1, y_2, \ldots, y_m)$ its translation, the NMT model maximizes $p(q_{\text{REF}}|q_{\text{SRC}})$ over all pairs $(q_{\text{SRC}}, q_{\text{REF}}) \in \mathcal{D}$. We build a large corpora $\mathcal{D} = \mathcal{D}_h \bigcup \mathcal{D}_e \bigcup \mathcal{D}_b \bigcup \mathcal{D}_t \bigcup \mathcal{D}_o$ with

Table 3: Performance of generic models on the test set.

Model	# params.	BLEU	CHRF	BERTscore	Time V100 GPU, bs=16
OpusMT [31]	77.9M	38.97	67.81	56.18	8m16s
m2m100_418m [8]	418M	23.07	58.23	36.51	1h32m13s
BergamotLarge [3]	108.4M	**41.35**	**74.95**	**61.09**	12m25s
Tiny.Untied	**16.9M**	33.5	62.12	62.6	**3m16s**

product-oriented and out-of-domain translations as follows. Table 5 shows statistics of the corpora.

Human-driven Corpus ($\mathcal{D}_h$). We actively maintain a human-driven dictionary with local terms from regions across Latin America (see **P6**). We manually correct the most critical queries for our business, overwriting the other partial corpora of $\mathcal{D}$.

Query-to-Item-driven Corpus ($\mathcal{D}_e$). Captures non-translatable tokens in highly engaged items; *i.e.*, items with high click-through rates after querying in Spanish. We consider both structured and unstructured data such as *item title*, *brand*, *product line*, *model* and *sport team* (see **P5**). The non-translatable tokens are inferred from the intersection of the tokens in these fields with those in the query itself. The translation is obtained with the Google API by masking the non-translatables. We have empirically found that the mask ADJECTIVE*i* ("ADJECTIVE" + identifier) gives us the best translations; *e.g.*, the query "*papel toalla viva*" is masked as "*papel toalla ADJECTIVE0*" and translated to "*paper towel ADJECTIVE0*", resulting in the final translation "*paper towel viva*".

Back-translation-driven Corpus ($\mathcal{D}_b$). We observe that our engagement-driven corpus can only go so far due to *cold start*, *presentation bias* and *low engagement* in tail Spanish queries. We address this limitation applying the *back-translation* technique [7, 30] to our extensive Item Catalog. We include titles that contain *non-translatable* tokens (see **P5**), as well as counter-example titles where those same tokens must be translated, thus gaining contextual information. For example, "*Costume with Crown*" is back-translated to "*disfraz con corona*" where *corona* is not the beer brand.

Top-queries Corpus ($\mathcal{D}_t$). We also include the most important queries of our e-commerce platform. This set of queries contains English and Spanish queries that should be processed differently (see Table 5). For Spanish queries we use the same strategy as in $\mathcal{D}_e$ and for English ones we use the same as in $\mathcal{D}_b$.

Out-of-Domain Corpus ($\mathcal{D}_o$). To address overfitting, we also include out-of-domain data, converting $\mathcal{D}$ in a rich mix of in-domain and out-of-domain useful for fine-tuning. We get the out-of-domain data from benchmarks used by OPUS-MT project [11].

5.2 Offline Evaluation

Our train and evaluation setting for the domain-adapted NMT model is as follows:

Dataset. Table 5 shows the composition of our Train-Validation-Test split of $\mathcal{D}$ at 70%:20%:10%. Random sampling ensures that the average length for both queries and their translations is similar across all splits.

Model training. We rely on the Marian toolkit [15] to train our model. First, we initialize Tiny.Untied with the parameters learned [2] on the WMT13 Spanish-English task [4] and then, we re-train to learn domain-specific parameters. We minimize the mean

Table 4: Six representative examples, which cover the predictions of each proposed component.

Component	Language Detection	P2 Example	P3 Example	P4 Example	P5 Example	P6 Example
Input Query	*strawberry cake*	*cake de fresa*	*pan / calabaza*	*gran turismo*	*cerveza corona*	*cabritas*
Translation Memory	not-in	not-in	in	not-in	not-in	not-in
Spanglish Detection	False	True	–	True	True	True
Spanglish Translatability	–	OptOut	–	OptIn	OptOut	OptOut
NMT Translation	–	*strawberry cake*	–	*grand tourer*	*corona beer*	*popcorn*
Experience	Ignore	automatic	Ambiguous	on-demand	automatic	automatic

Table 5: Corpus details. The average length, vocabulary size and examples of queries and translations.

Corp.	Size	Queries (Spanglish)			Translations (English)		
		Avg. len.	Vocab.	Example	Avg. len.	Vocab	Example
$\mathcal{D}_h$	9,545	2.38	2,808	fruta bomba	1.85	2,151	papaya
$\mathcal{D}_e$	515K	3.82	78,699	muñeca barbie	3.63	72,747	barbie doll
$\mathcal{D}_b$	3M	3.54	113,197	tortillas de maiz guerrero	3.03	107,246	guerrero corn tortillas
$\mathcal{D}_t$	740K	4.01	69,263	–	3.25	69,793	–
-Spanish	39K	2.22	17,973	taquitos	2.19	17,908	tacos
-English	700K	4.08	69,263	donas	3.29	63,795	donuts
$\mathcal{D}_o$	306K	8.97	74,730	visitar museos unicos	8.17	66,295	visit unique museums
$\mathcal{D}$	**4.7M**	**4.53**	**284,584**	–	**3.79**	**241,617**	–
-train (.7)	3.3M	**4.53**	246,915	–	**3.79**	210,873	–
-valid (.2)	958K	**4.53**	144,982	–	**3.79**	125,766	–
-test (.1)	480K	**4.53**	105,212	–	**3.79**	91,181	–

Table 6: Ablation study on the test set of our parallel corpora. Results by removing only one aspect of the training process.

Method	BLEU	CHRF	BERTscore	Translation for *pico de gallo en polvo* (GT: *pico de gallo powder*)
mixed-fine-tuning (ours)	**73.82**	**87.28**	**90.63**	pico de gallo powder
w/o $\mathcal{D}_h$	72.06	86.56	88.57	cock powder puff
w/o $\mathcal{D}_e$	70.66	85.15	86.52	pico de gallo powder
w/o $\mathcal{D}_b$	50.82	75.8	74.45	rooster pico powder
w/o $\mathcal{D}_t$	70.87	85.62	87.21	powdered rooster peak
w/o $\mathcal{D}_o$	68.78	83.26	85.88	powdered rooster beak
w/o *exp-smoothing*	67.04	83.88	88.36	pico de gallo powder
pre-trained	33.5	62.12	62.6	peak of rooster powder
from-scratch	70.85	85.58	87.51	pico de gallo powder

cross-entropy using the Adam optimizer [19] with default parameters, constant decaying factor and early stopping.

Model evaluation. At validation time, the model generates length-normalized vector translations using greedy beam search [17] and smoothed parameters with *exponential-smoothing* empirically set to $1e-4$, which performs better than no smoothing at all (see Table 6). Then, we obtain q_{HYP} truncating on $|q_{\text{HYP}}| \leq 2|q_{\text{REF}}|$. For comparison, we report results on three popular evaluation metrics: the n-gram-based *BLEU*, its character-based variant *CHRF*, and the transformer-based *BERTScore*.

5.2.1 NMT Ablation Study. Table 6 shows the results on our test set of eight ablated experiments that we performed as follows.

pre-trained. In this experiment, we omit the fine-tuning stage, which negatively affects all the metrics, causing a drop of 62% in BLEU. The Example column shows that fine-tuning on $\mathcal{D}$ helps the NMT model detect non-translatable tokens and choose more appropriate translations for polysemic words in general.

from-scratch. To investigate the impact of fine-tuning on the performance of the pre-trained NMT model, we experimented with training a model from scratch with the same architecture. The results show that the model trained from scratch achieves a BLEU score of 70.85, while the fine-tuned model achieved a score of 73.82. This suggests that initializing the model with pre-trainied weights is effective in improving the quality of the translations.

w/o $\mathcal{D}_h$, w/o $\mathcal{D}_e$, w/o $\mathcal{D}_b$, w/o $\mathcal{D}_t$, and w/o $\mathcal{D}_o$. These ablated experiments exhibit the individual contribution of each corpus when removed so that the model is trained on the remaining four. Our model achieves its best performance when all of the corpora are included in the training set, proving the effectiveness of the proposed methodology. Moreover, from a qualitative perspective, the most-right column in Table 6 illustrates the effect of each corpus (except $\mathcal{D}_e$) in keeping *pico de gallo* unaltered. Due to its diverse vocabulary and its large covering of non-translatable tokens, the corpus that contributes the most is $\mathcal{D}_b$, followed by $\mathcal{D}_o$.

We also tested the model by changing the default Marian-NMT value of only one hyperparameter. Training on the same corpora $\mathcal{D}$, we varied the clip gradient normalization to *clip-norm* = 1, the optimization function to *cost-type* = ce-sum, and the exponent of the translation score normalization to *normalize* = 0. All these changes cause a drop in the metrics but they are not as marked as that of *exponential-smoothing* = 0.

6 IMPACT OF THE CLIR EXPERIENCE

In Table 4 we show queries mentioned throughout the document as they flow through the components depicted in Figure 1.

We measure the impact in relevance using Normalized Discounted Cumulative Gain (nDCG) [13]. The input sample of 300 queries was taken at random without replacement from the Spanish traffic in the last 12 months. Then the ranked items resulting from the original Spanish query were assigned to CONTROL and the ones from the translated query to TREATMENT. Each *query-item* pair was manually graded with a 5 point base, with 0 being Irrelevant and 4 being Very Relevant. Grading required bilinguality to understand the original intention in Spanish and accurately asses the items described in English. A fixed results size was then enforced for each query and minimum grades assigned to the missing items. Finally, nDCG@5 reported a +70% relevance gain and nDCG@10 a +73%, both with p-Value < 0.05.

Beyond relevance, we measured impact through an A/B test where the Spanish traffic from 33% of customers went to CONTROL (*i.e.*, without our CLIR) and the Spanish traffic from another 33% got the CLIR experience. The null hypothesis was rejected with statistically significant lifts in Spanish GMV, orders and first time buyers. The test also reported a 20% lift in Spanish query impressions.

7 CONCLUSIONS

In this paper, we presented a query translation experience based on query translatability and mixed fine-tuning that achieves significant progress on the following aspects:

- To the best of our knowledge, this work is the first to tackle the decision of translating a query from a holistic perspective that goes beyond language identification and predicts whether or not the search intent requires translation to be optimally full-filled. A/B testing showed a significant lift in business metrics for this strategy.
- To the best of our knowledge, this is the first approach to combine *catalog back-translations*, *query engagement analysis*, and *general-domain translations* into an e-commerce-specific corpus which supports a scalable solution, detection of non-translatables and higher search relevance, as shown by nDCG results.
- Mixed-domain fine-tuning has the advantage of exposing the model to more data, and this improves the fluency of the translations. Additionally, it can also help to keep the model from overfitting to the specific domain data, which can be useful when working with small datasets.
- Allowed by the short nature of queries and the adaptation opportunity to our specific e-commerce domain, a fast and generic encoder-decoder architecture for translation has been tuned. The low-latency of the resulting model can serve an ever growing Hispanic community.

As future work we will study the differences in conversion for the same search intentions when approached from different languages. This will provide the means for contextualization and further improvement of the translations.

REFERENCES

[1] Aman Ahuja, Nikhil Rao, Sumeet Katariya, Karthik Subbian, and Chandan K. Reddy. 2020. Language-agnostic representation learning for product search on e-commerce platforms. In *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining*. Association for Computing Machinery, Inc, 7–15. https://doi.org/10.1145/3336191.3371852

[2] Bergamot. 2022. NMT models for Bergamot. https://github.com/browsermt/students

[3] Nikolay Bogoychev, Roman Grundkiewicz, Alham Fikri Aji, Maximiliana Behnke, Kenneth Heafield, Sidharth Kashyap, Emmanouil-Ioannis Farsarakis, and Mateusz Chudyk. 2020. Edinburgh's Submissions to the 2020 Machine Translation Efficiency Task. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*. Association for Computational Linguistics, Online, 218–224. https://doi.org/10.18653/v1/2020.ngt-1.26

[4] Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, 1–44. https://aclanthology.org/W13-2201

[5] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[6] Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. Enabling Code-Mixed Translation: Parallel Corpus Creation and MT Augmentation Approach. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*. 131–140.

[7] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. Association for Computational Linguistics, 489–500. https://doi.org/10.18653/V1/D18-1045

[8] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation.

[9] Antonio Flores. 2020. 2015, Hispanic population in the United States statistical portrait. https://www.pewresearch.org/hispanic/2017/09/18/2015-statistical-information-on-hispanics-in-united-states/

[10] Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. CoMeT: Towards Code-Mixed Translation Using Parallel Monolingual Sentences. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*.

[11] Helsinki-NLP. 2022. opus-mt-tc-big-cat-oci-spa-en. https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-cat_oci_spa-en

[12] Qie Hu, Hsiang-Fu Yu, Vishnu Narayanan, Ivan Davchev, Rahul Bhagat, and Inderjit S. Dhillon. 2020. Query transformation for multi-lingual product search. In *SIGIR 2020 Workshop on eCommerce*.

[13] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (10 2002), 422–446. https://doi.org/10.1145/582415.582418

[14] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification - ACL Anthology. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain. https://aclanthology.org/E17-2068/

[15] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F.T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*. Association for Computational Linguistics (ACL), 116–121. https://doi.org/10.18653/V1/P18-4020

[16] Heafield Kenneth Junczys-Dowmunt Marcin, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective High-Quality Neural Machine Translation in C++. *CoRR* abs/1805.12096 (2018).

[17] Yoon Kim and Alexander M Rush. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1317–1327. https://doi.org/10.18653/v1/D16-1139

[18] Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From Research to Production and Back: Ludicrously Fast Neural Machine Translation. *EMNLP-IJCNLP 2019 - Proceedings of the 3rd Workshop on Neural Generation and Translation* (2019), 280–288. https://doi.org/10.18653/V1/D19-5632

[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/ARXIV.1412.6980

[20] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada, 67–72. https://aclanthology.org/P17-4012/

[21] Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Carl Case, and Paulius Micikevicius. 2018. OpenSeq2Seq: Extensible Toolkit for Distributed and Mixed Precision Training of Sequence-to-Sequence Models. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics (ACL), 41–46. https://doi.org/10.18653/V1/W18-2507

[22] Mandar Kulkarni, Soumya Chennabasavaraj, and Nikesh Garera. 2022. Study of Encoder-Decoder Architectures for Code-Mix Search Query Translation. https://doi.org/10.48550/ARXIV.2208.03713

[23] Robert Lim, Kenneth Heafield, Hieu Hoang, Mark Briers, and Allen Malony. 2018. Exploring Hyper-Parameter Optimization for Neural Machine Translation on GPU Architectures. *CoRR* (5 2018).

[24] Marius Lindauer, Katharina Eggensperger, Matthias Feurer, André Biedenkapp, Difan Deng, Carolin Benjamins, Tim Ruhkopf, René Sass, and Frank Hutter. 2022. SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization. *Journal of Machine Learning Research* 23, 54 (2022), 1–9. http://jmlr.org/papers/v23/21-0888.html

[25] Hanqing Lu, Youna Hu, Tong Zhao, Tony Wu, Yiwei Song, and Bing Yin. 2021. Graph-based Multilingual Product Retrieval in E-Commerce Search. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*. Association for Computational Linguistics, Stroudsburg, PA, USA, 146–153. https://doi.org/10.18653/v1/2021.naacl-industry.19

[26] Sourab Mangrulkar, Amazon Bengaluru, India M Ankith S, India Vivek Sembium, and Ankith M S. 2022. Multilingual Semantic Sourcing using Product Images for Cross-lingual Alignment. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion)*, Vol. 1. ACM, 11. https://doi.org/10.1145/XXXXXX.XXXXXX

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *… of the 40Th Annual Meeting on …* July (2002), 311–318. https://doi.org/10.3115/1073083.1073135

[28] Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *10th Workshop on Statistical Machine Translation, WMT 2015 at the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015 - Proceedings*. Association for Computational Linguistics (ACL), 392–395. https://doi.org/10.18653/V1/W15-3049

[29] Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018. Word Embeddings for Code-Mixed Language Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

[30] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, Vol. 1. Association for Computational Linguistics (ACL), 86–96. https://doi.org/10.18653/V1/P16-1009

[31] Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*. Lisbon, Portugal.

[32] Kirti Vashee. 2022. The impact of MT on the Global Ecommerce Opportunity. https://blog.modernmt.com/the-impact-of-mt-on-the-global-ecommerce-opportunity/

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. 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/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[34] Liang Yao, Baosong Yang, Haibo Zhang, Weihua Luo, and Boxing Chen. 2020. Exploiting Neural Query Translation into Cross Lingual Information Retrieval. In *SIGIR eCom 2020*. https://doi.org/10.48550/arxiv.2010.13659

[35] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

[36] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, Jimmy Lin, and David R Cheriton. 2022. Towards Best Practices for Training Multilingual Dense Retrieval Models. (4 2022). https://doi.org/10.48550/arxiv.2204.02363