Cross-Lingual Product Retrieval in E-Commerce Search

Wenya Zhu[1](✉), Xiaoyu Lv[1], Baosong Yang[1], Yinghua Zhang[2], Xu Yong[1],
Linlong Xu[1], Yinfu Feng[1], Haibo Zhang[1], Qing Da[1], Anxiang Zeng[3],
and Ronghua Chen[4]

[1] Alibaba (China) Technology Co., Ltd., Hangzhou, China
`zhuwenya1991@gmail.com`, `{linlong.xll,yinfu.fyf,daqing.dq}@alibaba-inc.com`
[2] Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
`yzhangdx@cse.ust.hk`
[3] Nanyang Technological University, Singapore, Singapore
`zeng0118@ntu.edu.sg`
[4] Fudan University, Shanghai, China
`chenrh@fudan.edu.cn`

Abstract. Cross-lingual product retrieval (CLPR) recalls semantically relevant products that match multilingual search queries. It plays a crucial role in E-commerce sites to serve cross-border customers. However, there exists no public large-scale dataset on CLPR, hindering the research on this topic. We present CLPR-9M (https://tianchi.aliyun.com/dataset/dataDetail?dataId=121505), the first large-scale CLPR dataset containing 9 million query-product pairs, covering 10 major commodity categories and 3 language pairs, mined from real-world user logs. We also release a test dataset, annotated by bilingual experts with fine-grained labels. We build our baselines upon the widely used cross-lingual embedding retrieval framework and improve it from a range of aspects, including the pretrain-finetune paradigm, negative sampling, as well as optimization objective. Benchmarks are assessed and reported using multiple evaluation metrics, and will be beneficial for future research in this area.

Keywords: Cross-lingual information retrieval · E-commerce search

1 Introduction

With the growth of international market, E-commerce websites have to cope with not only monolingual but also multilingual queries, in order to serve cross-border customers. For example, a seller from America can serve customers from Southeastern Asia. In this case, the product information is written in English, while the query may be in Thai, Filipino, or Bahasa Indonesia. The products remain monolingual for two reasons. Firstly, it requires non-trivial efforts for sellers to provide multilingual item descriptions; Secondly, building the multilingual item indexes with the machine translation is limited by the quality of the machine translation. We refer to the product retrieval [12,27] in this setting

J. Gama et al. (Eds.): PAKDD 2022, LNAI 13281, pp. 458–471, 2022.
https://doi.org/10.1007/978-3-031-05936-0_36

as cross-lingual product retrieval (CLPR), where the product descriptions and the user queries are in different languages. As more sellers are expanding their business in emerging markets, the CLPR setting is becoming popular.

However, few studies explored CLPR, due to the lack of in-domain dataset, especially for the state-of-the-art deep learning models which heavily depends on large-scale training samples. Although [28] and [17] paid their attention to out-of-domain cross-lingual retrieval tasks, these studies may fail to generalize to the E-commerce domain due to non-trivial domain discrepancy. Figure 1 provides the taxonomy of information retrieval datasets from the domain and the language aspects.

To fill the gap, we collect and release the first large-scale cross-lingual product retrieval dataset (CLPR-9M). We construct the training set by extracting query-product pairs from real-world user logs. Since labeling the negative samples requires non-trivial efforts, past studies obtained the negatives by sampling with the human-crafted strategy, which has achieved reasonable performance [6,7,21,27]. In our dataset, we provide the irrelevant query-product pairs from two sampling strategies, including random sampling and category-based sampling. In total, the training set is composed of 9 million query-product relevant pairs that are from 10 categories. The queries are in Russian, Spanish, and English, while the product titles are in English only. To evaluate the generalization ability of the retrieval model, we provide the high-quality test set with three labels (relevant, weak relevant, and irrelevant) by carefully manual annotation. As shown in Fig. 2, we provide several samples from the proposed dataset.

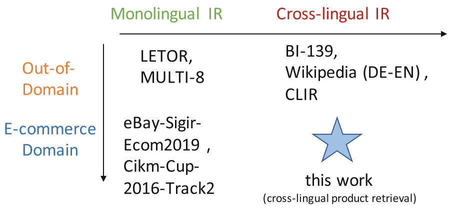

Monolingual IR
Cross-lingual IR
Out-of-Domain
E-commerce Domain
LETOR, MULTI-8
BI-139, Wikipedia (DE-EN), CLIR
eBay-Sigir-Ecom2019 , Cikm-Cup-2016-Track2
this work
(cross-lingual product retrieval)

Fig. 1. Taxonomy of information retrieval (IR) datasets. We divided the information retrieval into monolingual and cross-lingual settings. Our work in this paper is to provide benchmarks for the cross-lingual IR in E-commerce domain. LETOR [15] and MULTI-8 [22] are the monolingual IR datasets in Wikipedia. Wikipedia (DE-EN) [19], CLIR [18] and BI-139 [22] are the datasets for cross-lingual IR in Wikipedia.

Building cross-lingual retrieval models has its unique challenges, such as how to bridge the lexical gap between languages [14]. Recently, the pretrained language models, such as multilingual BERT (M-BERT) [5] and XLM [11] can induce shared cross-lingual semantic space by learning the pretrained tasks based on sentence-aligned parallel data. We finetune the pretrained cross-lingual language model on the dataset, and provide extensive experiments to explore the loss function and negative sampling strategy. For the loss function, we propose a bi-log loss that maximizes the log-likelihood of positives from query and item directions. For the negative sampling, we compare random sampling, category-based sampling, and mixes of them. The experiments show that random negative sampling with bi-log loss can achieve a decent performance.

2 Related Work

2.1 Product Retrieval Datasets

The monolingual datasets (e.g., CIKM Cup 2016 Track 2 [1] and eBay Sigir Ecom 2019 [2]) have enabled the development of product retrieval for E-commerce search. However, to the best of our knowledge, there is no public large-scale dataset for the cross-lingual product retrieval task. A similar effort is cross-lingual information retrieval datasets for general domains, such as Wikipedia (DE-EN) [19], CLIR [18], and BI-139 [22]. They may not be applied to the E-commerce domain due to the non-trivial domain discrepancy. In addition, the multilingual queries of these datasets are extracted from the title or the first sentence of the document, rather than real-world user inputs. The relevance label is determined by various hand-crafted rules, such as smoothing out the BM25 score into discrete relevance labels in [22]. Our contribution differs from the above studies in three aspects: 1) CLPR-9M is the first large-scale dataset for cross-lingual information retrieval in E-Commerce search; 2) All multilingual search queries are from real users, and the dataset is closer to the real-world application; 3) CLPR-9M provides a high-quality benchmark by human annotation with finer-grained levels of relevance.

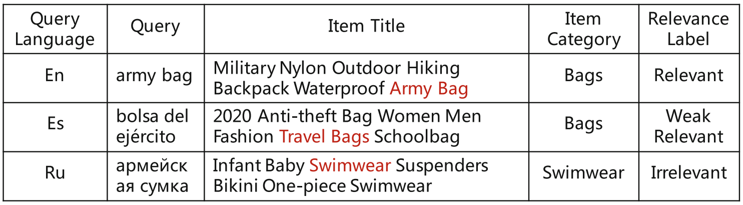
2.2 Product Retrieval Methods

With the success of deep learning, a large number of neural network based models have been proposed to enhance traditional product retrieval methods (e.g., BM25[16], LSI [4]) and learning to rank methods [10]. The neural retrieval models represent queries and products as dense vectors, which are further exploited to produce relevance scores. Particularly, DSSM [7] and its variant CDSSM [21] have pioneered the context of using deep neural networks for relevance scoring. Van Gysel et al. [23] proposed a latent vector space model (LSE) to learn the query and product representations with the entities as bridge. Zhang et al. [27] proposed two tower model to achieve the personalized and semantic retrieval goal. These methods have shown promising results on monolingual product retrieval tasks.

Nevertheless, due to the lack of large-scale public datasets, few studies in terms of deep model explore the cross-lingual scenario, especially for the E-commerce domain. Existing cross-lingual information retrieval (CLIR) systems usually adopt a translation-based approach that consists of three stages, including language identification, machine translation, as well as monolingual information retrieval [3,13,29]. However, the performance of the translation-based approaches is limited by the quality of the language identification and machine translation [29]. Recently, multiple pretrained language models have been developed, such as M-BERT [5] and XLM [11], that model the underlying data distribution and learn the linguistic patterns or features across languages and have been applied in cross-lingual information retrieval [9]. In this way, the cross-lingual information retrieval can be trained end-to-end, thus avoids the error propagation from language identification and machine translation.

3 Dataset

In this section, we introduce the construction of the CLPR-9M dataset. The dataset is composed of a set of query-product triplets $\{\mathbf{q}_n, \mathbf{i}_n, r_n\}_{n=1}^{N}$, where $\mathbf{q}_n$, $\mathbf{i}_n$, and r_n denote the query, product, and the semantic relevance between the query and the product, respectively. The query $\mathbf{q}_n$ is a sequence of m words, and there is $\mathbf{q}_n = \{q_{1,n}, q_{2,n}, \cdots, q_{m,n}\}$. Similarly, the product $\mathbf{i}_n$, which is also a sequence of k words, is denoted by $\mathbf{i}_n = \{i_{1,n}, i_{2,n}, \cdots, i_{k,n}\}$. There are 3 possible values for the relevance with 0 to represent irrelevant, 1 to represent weak relevant and 2 to represent relevant. In the following, we describe the collection of the training set and the annotation of the test set, followed by a brief summary of dataset statistics.



Query Language	Query	Item Title	Item Category	Relevance Label
En	army bag	Military Nylon Outdoor Hiking Backpack Waterproof Army Bag	Bags	Relevant
Es	bolsa del ejército	2020 Anti-theft Bag Women Men Fashion Travel Bags Schoolbag	Bags	Weak Relevant
Ru	армейская сумка	Infant Baby Swimwear Suspenders Bikini One-piece Swimwear	Swimwear	Irrelevant

Fig. 2. The samples of the dataset CLPR-9M. For each query-product pair, we provide the query language, the query content, the item title, the item category and relevance label. The terms in item title denoted what the product is are marked in red. (Color figure online)

3.1 Training Data Mining

The training set is mined from real-world user logs. Since it is difficult to determine whether a query-product pair is weakly related from the user logs, the relevance is a binary value in the training set. A semantically relevant query-product pair is considered as a positive sample, and similarly, an irrelevant pair is negative. Recent studies [6,27] suggest that using click results as positives and randomly sampling negatives can provide a reasonable model performance. Inspired by the previous studies, we randomly sample clicked pairs of 10 categories from online 1-month logs as the positives. The category of the query-product pair is determined by that of the product. There are several sampling strategies to obtain negative pairs. We provide negatives with two sampling strategies: **Random sampling** (for each query, we randomly sample products from all candidate products as negatives) and **Category-based sampling** (for a positive pair $\{\mathbf{q}_n, \mathbf{i}_n\}$, randomly sample products under the same domain of $\mathbf{i}_n$ as irrelevant products). Compared with random sampling, category-based sampling can produces hard negatives, since the products under the same domain tend to be similar.

3.2 Human Annotated Test Set

To build the test set, we first select the query-product pairs clicked by users as seed positive samples. Then, for a query in an arbitrary seed positive pair, we obtain the potential irrelevant products with three sampling strategies, and hence form potential negative pairs. To include more hard negatives, we added **Unclicked Impressed Sampling**: random sampling the products impressed to the user but not clicked, except for **Random sampling** and **Category-based sampling**. Notice that we do not utilize **Unclicked Impressed Sampling** to form the negatives in the training data, since the impressed products usually have some degree of relevance with search queries and the users do not click them may due to personal preference. Finally, each query-product pair is rated by two bilingual experts with three labels, namely "relevant", "weak relevant", and "irrelevant". The annotation instruction is provided in the Appendix. A pair with same labels from two bilingual experts is accepted; otherwise, the third language expert will make a decision.

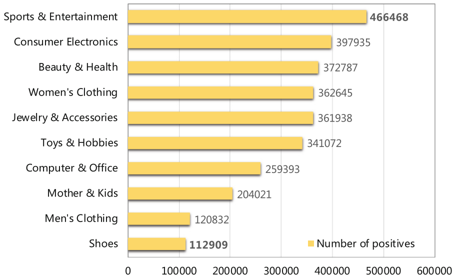


Fig. 3. The number of positive samples per category in the training set.

3.3 Dataset Statistics

The dataset contains 10 categories, and a total of 9 million query-product pairs for training, 21, 700 pairs for testing. The training dataset contains 3 million relevant query-product pairs, 3 million irrelevant query-product pairs from random sampling, and 3 million irrelevant query-product pairs from category-based sampling. Figure 3 shows the number of query-product pairs for each category in the training set. In both the training and test set, the product title is in English, and the search queries have 3 languages, namely English(En), Russian(Ru) and Spanish(Es). Table 1 shows the statistics per language of the dataset. If all the tokens of a query appear in the product title, the query-product pair is considered as an "exact-match" pair; otherwise, it is an "inexact-match" pair. The inexact-match pairs cannot be handled by simple template matching methods, making the CLPR task challenging. The number of inexact-match pairs is 2.44 times larger than that of exact-match pairs, which indicates that CLPR on the CLPR-9M is indeed a demanding task.

Table 1. The statistics for the training and the test dataset. For each language X, we show the total number of queries (#Query) in language X, and the number of query product pairs (#QP pairs) where the query in language X. The number of products is shown in column #Product.

Dataset	#Query			#QP pairs			#Product
	English	Russian	Spanish	English	Russian	Spanish	
Training	1.14M	0.647M	0.648M	4.26M	2.36M	2.38M	1.70M
Test	0.76K	0.73K	0.68K	7.6K	7.35K	6.8K	26K

3.4 Human Evaluation on Dataset Quality

To evaluate the quality of the training data, we hire the bilingual expert to evaluate the relevance of 20,000 query product pairs randomly selected from the training data. For the positives in the training data, the accuracy of the labels is 80%. The error rate of the negatives is 0.02% and 2% for random sampling and category-based sampling, respectively. Besides, the agreement rate of two raters for test data annotation is 96.1%.

4 Baseline Approaches

In this section, we present a neural network retrieval model as a baseline model for the CLPR task on the CLPR-9M dataset. Motivated by the framework in [6], the model first converts the query and the product tokens into embeddings, and then generates a relevance score based on the extracted embeddings. An overview of the model is shown in Fig. 4. In the following, we first describe the method to extract embeddings and the scoring function to measure relevance. Then we explore two design choices, namely negative sampling strategies and loss functions.

4.1 Retrieval Model

As shown in 4, there are two major components in the retrieval model, namely the embedding model that encodes the query and product tokens into dense vectors and the scoring function that measures the relevance of the query-product pair. Both the query encoder and the product encoder adopt the same multi-layer transformer architecture, and the parameters are shared. With the self-attention mechanism [24], the transformer-based encoder outputs context-based token embeddings. The query embedding $\vec{q_n}$ and product embedding $\vec{i_n}$ are obtained by the average pooling of the token embeddings. The encoders are initialized with the pre-trained cross-lingual language model, such as M-BERT [5] and XLM [11]. However, existing public pretrained language models are trained on the Wikipedia corpus, which may not generalize to the E-commerce domain. To avoid the domain discrepancy, we utilize the E-commerce corpus to learn the cross-lingual language model, denoted as EXLM, with the pretrained task proposed in XLM [11]. In detail, the Translation Language Modeling is trained with translated query pairs, and the Masked Language Modeling is trained with monolingual queries and English item titles.

After obtaining the query and item embeddings, we choose cosine similarity as the score function $S(q_n, i_n)$ which is commonly used in the retrieval task [6]:

$$S(q_n, i_n) = \frac{\vec{q_n} \cdot \vec{i_n}}{\|\vec{q_n}\| \|\vec{i_n}\|}, \tag{1}$$

where $\cdot$ denotes the dot-product of two vectors and $\| \cdot \|$ is the l_2-norm of the vector.

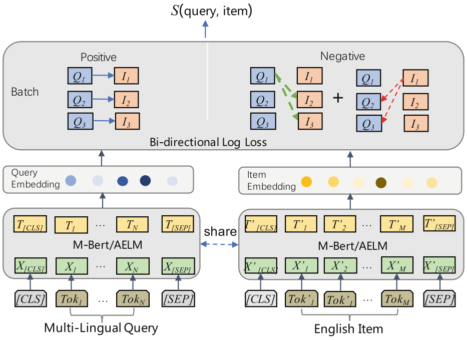


Fig. 4. The model architecture for CLPR with the batch negatives. The query embedding and item embedding is obtained by average pooling of outputs of multi-layer transformers. The transformers are initialized by the pre-trained cross-lingual model (M-BERT, XLM etc.). The solid line in the batch denotes the positive pairs. The dotted line denotes the batch negatives obtained in two ways. One is combining the irrelevant items with the query (the green dotted line), and the other is combining the irrelevant queries with the item (the red dotted line). (Color figure online)

4.2 Negative Sampling

Labeling negative samples for the retrieval task requires a large amount of labor and time cost. In the past studies, negatives are usually obtained by sampling based on human-crated rules. Here, we compared several sampling strategies, including random sampling, category-based sampling and a mixing strategy that combines random sampling and category-based sampling.

- **Random sampling:** for each query, we randomly sample items from all candidate items as negatives.
- **Category-based sampling:** for a positive pair $\{q_n, i_n\}$, randomly sample items under the same domain of i_n as irrelevant items.
- **Mixing strategies:** compared with random sampling, category-based sampling can produces hard negatives, since the items under the same domain tend to be similar. We explore two ways to combine the two sampling strategies. One is to train the model with random sampling first and then with category-based sampling (Random -> Category). The other is to train the model with category-based sampling first and then train with random sampling (Category -> Random).

Both random sampling and category-based sampling are computational expensive, with computational complexity $O(N_q \times N_i)$, where N_q is the number of queries and N_i is the candidate item pool size. To reduce computational complexity, the **batch negative** is adopted to approximate random sampling, where the irrelevant items for the query are the positive items from other queries in the same batch, as shown in Fig. 4. To implement category-based sampling with batch negatives, we organize the positive pair with the same category together.

4.3 Loss Function

We consider two popular loss functions for the retrieval task, namely the triplet loss [20,25] and the log-likelihood loss [21,26]. The triplet loss enforces a positive pair, denoted by $\{q_n, i_n^+\}$, to separate from a negative pair, denoted by $\{q_n, i_n^-\}$, by a distance margin m and is defined as:

$$\mathcal{L}_{triplet} = \sum_{n=1}^{N} max(0, D(q_n, i_n^+) - D(q_n, i_n^-) + m), \tag{2}$$

where $D(u, v)$ is a distance metric between vectors u and v, and is defined as $1 - S(u, v)$ in this paper.

The log-likelihood objective with the softmax function aims to place positives over the negatives. For the positive pair $\{q_n, i_n^+\}$, we can utilize the irrelevant item i_n^- for query q_n to compose the negative sample $\{q_n, i_n^-\}$ or the irrelevant query q_n^- for item i_n to compose the negative sample $\{q_n^-, i_n\}$. Thus, we can compute two log loss, one with $\{q_n, i_n^-\}$ as negatives (denotes as q-log loss) and the other with $\{q_n^-, i_n\}$ as negatives (denoted as i-log loss). The q-log loss and i-log loss are defined as:

$$\mathcal{L}_{q\_log} = -\frac{1}{N} \sum_{n=1}^{N} \log \frac{exp(S(q_n, i_n))}{exp(S(q_n, i_n)) + \sum_{i_k^- \in I_{q_n}} exp(S(q_n, i_k^-))}, \tag{3}$$

$$\mathcal{L}_{i\_log} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{exp(S(q_n, i_n))}{exp(S(q_n, i_n)) + \sum_{q_k^- \in I_{i_n}} exp(S(q_k^-, i_n))}, \tag{4}$$

where I_{q_n} is the set of irrelevant items for the query q_n, and I_{i_n} is the set of irrelevant queries for the item i_n. The sum of $\mathcal{L}_{q-log}$ and $\mathcal{L}_{i\_log}$ is denoted as bi-log loss $\mathcal{L}_{bi\_log}$. Figure 4 illustrates the bi-log loss computed with batch negatives.

$$\mathcal{L}_{bi\_log} = \mathcal{L}_{q\_log} + \mathcal{L}_{i\_log}. \tag{5}$$

5 Experiments

5.1 Evaluation Metrics

AUC is widely used to evaluate the product retrieval system. However, it cannot measure the effectiveness of an individual query, since it is computed over the whole test set. Inspired by the **Group AUC (GAUC)** proposed in [30], we define **GAUC** for the retrieval task as the mean of the AUC for each query. Both **AUC** and **GAUC** can only measure the ability to distinguish relevant and irrelevant pairs. To measure the ability to distinguish relevant, weak relevant, and irrelevant query-product pairs, we utilize **NDCG** [8], which is a popular metric for the ranking algorithms. In detail, the NDCG computes the similarity between the ranking results for each query and that based on relevance labels, and then is averaged over all test queries. Notice that the weak relevant label is used as irrelevant label when computing the **AUC** and **GAUC** metric.

Table 2. AUC of different negative sampling strategies on the CLPR task.

Model	Ru⇒En	Es⇒En	En⇒En	AVG
Category	81.47	77.36	82.76	80.53
Random	82.12	76.84	82.51	80.49
Category -> Random	82.05	77.30	82.49	80.61
Random -> Category	**82.74**	**77.39**	**83.19**	**81.11**

5.2 Experimental Setting

The query encoder and item encoder are initialized with the pretrained cross-lingual language model, the 12-layer transformers with 768 hidden size. The max sequence length of the query encoder and the item encoder is 20 and 40, respectively. To finetune the pretrained cross-lingual language model, we use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay of 0.01, and learning rate of 3×10^{-5}. The margin value in the triplet loss and bi-log loss is set to 0.2, which leads to the best performance. We train all models by 10 epoches with a batch size of 512.

5.3 Effect of Negative Sampling

Table 2 shows the performance of various negative sampling strategies, including category-based sampling (Category), random sampling (Random), transferring random sampling to category-based sampling (Random -> Category), and transferring category-based sampling to random sampling (Category -> Random). All sampling strategies are implemented with *batch negatives*. Although category-based sampling may produce hard negatives, random sampling exhibits better performance than category-based sampling. This shows that the presence of easy negatives in training data is necessary. Besides, mixing easy and hard negatives in the training process is advantageous. Our experiment shows that transferring easy to hard achieves better performance than transferring hard to easy negatives. Consequently, transferring random sampling to category-based sampling (Random -> Category) is applied as the default in subsequent experiments.

5.4 Effectiveness of Bi-Directional Log Loss

We compare two loss functions, the triplet loss and our proposed bi-log loss. Since the bi-log loss is the sum of q-log loss and i-log loss, we further analyse the effectiveness of these two losses, respectively. Table 3 shows the results obtained with various loss functions. The i-log loss achieves better performance than the q-log loss, indicating forming the negatives by sampling the irrelevant multilingual queries given the item is more effective in the CLPR task. Although using q-log and i-log loss independently cannot achieve better performance than using the triplet loss, bi-log loss performs the best in terms of the AUC metric on all language pairs. This observation suggests that the q-log loss and i-log loss are complementary to each other.

Table 3. AUC of various loss functions on the CLPR task.

Model	Ru⇒En	Es⇒En	En⇒En	AVG
Triplet	79.28	76.88	78.71	78.29
Q-Log-Loss	76.85	73.87	77.6	76.11
I-Log-Loss	76.95	74.47	79.56	76.99
Bi-Log-Loss	**81.47**	**77.36**	**81.47**	**80.10**

Table 4. AUC, GAUC and NDCG of different models on CLPR task

Model	Ru⇒En			Es⇒En			En⇒En			AVG		
	AUC	GAUC	NDCG	AUC	GAUC	NDCG	AUC	GAUC	NDCG	AUC	GAUC	NDCG
DSSM	79.08	84.04	91.40	76.00	83.13	91.80	81.63	86.22	93.04	78.90	84.46	92.08
M-Bert+FT	82.74	87.81	93.05	77.39	84.37	91.90	83.19	88.63	94.21	81.11	86.94	93.05
EXLM+FT	**83.78**	**88.76**	**94.43**	**78.43**	**84.62**	**92.03**	**83.35**	**88.71**	**94.37**	**81.85**	**87.36**	**93.61**

5.5 Main Results

Given the best negative sampling strategy and loss function explored in the above sections, we explore the model architectures and the pretrained cross-lingual language models in this section. The best performance is reported as the baseline of the CLPR-9M dataset. Table 4 shows the results in terms of various evaluation metrics (AUC, GAUC and NDCG). **M-Bert+FT** and **EXLM+FT** finetune the pretrained models M-Bert and EXLM respectively. The overall performance of the models by finetuning pretrained language models achieves better performance than **DSSM**. The performance of **EXLM+FT** is better than that of **M-Bert+FT**, which indicates that the pretraining with parallel corpora and in-domain data can facilitate the CLPR learning. For all models, the English-English language direction achieves the best performance. This suggest that cross-lingual training is more challenging than monolingual training. The best performance is reported as the baseline of the CLPR-9M dataset.

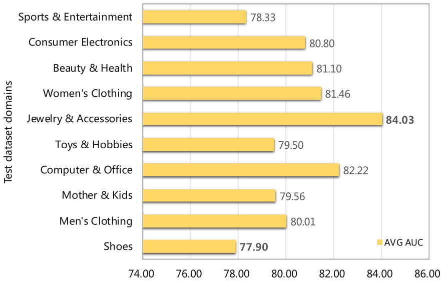


Fig. 5. AUC of XLM+Bilog on different categories.

5.6 Effect of the Category

Figure 5 illustrates the AUC on 10 categories for the best model. We find that the different categories have various levels of learning hardness. For example, the performance of the category **Sports** & **Entertainment** with the largest training data size ranked ninth place, while the category **Jewelry** & **Accessories** outperforms other categories with the medium training data size.

6 Conclusions

We construct CLPR-9M, a large-scale cross-lingual product retrieval benchmark. The CLPR-9M includes the training data by sampling online click logs, and the manually labelled test data. We conduct extensive experiments comparing different negative sampling strategies, and baseline models. Additionally, cross-lingual data facilitates the study of the cross-lingual language model.

Acknowledgement. We would like to thank for the support from the the National Key R&D Program of China under Grant 2018YFB1403200.

A Appendix

A.1 The annotation instructions for test dataset

The test set is obtained by the annotation of bilingual experts. We provide the detailed rating criteria to guarantee labeling Quality. For each label (relevant, weak relevant and irrelevant), we provide multiple criteria and the example to illustrate each criterion. The rating criteria and examples are shown in Table 5.

Table 5. The rating criteria and examples for human raters

Label	Criteria	Examples	
		Query	Item Title
Relevant	The item is consistent with the intention of query, and the title of the product exactly matches the literal or meaning of query	Wedding dress	**Wedding Dress** Long Sleeve Sheer Neck Appliques Bridal Gowns 2020 Spring
	The item is consistent with the intention of query, and item title does not match query literal, but it is synonym or abbreviation or original meaning	Mobile phone	Apple iPhone X 4G LTE **Mobile Cellphone** 3 GB RAM 64 GB 256 GB ROM 5.8”
	Brand and category have exactly the same intention as query	Apple iphone 12	**Apple iPhone 12** 5G LTE Mobile Phone 64 GB 256GB ROM 6.1”
	The item is consistent with the intention of query, but query is the hypernym	Lady shoes	Eilyken 2021 New Summer Fashion Design High heels **Ladies Sandals** Open Toe Shoes
Weak Relevant	The item is consistent with the intention of query, but query is the hyponym	Calf leather shoes	TUINANLE 2021 Autumn Winter **Shoes** Women Plush Snow Boot Heel Fashion Keep Warm Women’s Boots Woman Size 36–42 Ankle Botas Pink
	The main product of title is consistent with the main product of query, but the attributes are different	64G usb driver	20pcs/lot Hot sale USB Flash Drive pendrive 8 GB 16 GB 32 GB
	The item is accessory of query	iPhone 11	Camera Lens Protection **Phone Case For iPhone 11** 12 Pro Max 8 7 6 6s
Irrelevant	The brand for item is different with query	Huawei phone case	Flower **Case For Samsung Galaxy** A50 A51 Plus Ultra S10E TPU
	The category for item is different with query	Apple iPhone	**Apple IPad Mini** 1st 7.9” 2012 16 Gb Silver Black 80% New Original Refurbish
	Item is related to intention of the query, and both belong to the same concept/category/industry, but not the same kind of products	Slippers	Eilyken 2021 New Summer Fashion Design Weave Women **Sandals**
	The item is totally different with query	Power cable	15 Pack LED S14 Replacement **Light Bulbs**, Warm White Edison Bulbs for Outdoor String Lights

References

1. CIKM Cup 2016 Track 2 (2016). https://competitions.codalab.org/competitions/
2. eBay SIGIR 2019 eCommerce search challenge (2019). https://sigir-ecom.github.io/ecom2019/data-task.html
3. Chen, A., Gey, F.C.: Combining query translation and document translation in cross-language retrieval. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 108–121. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30222-3_10
4. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Huang, J.T., et al.: Embedding-based retrieval in Facebook search. In: KDD, pp. 2553–2561 (2020)
7. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: CIKM, pp. 2333–2338 (2013)
8. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. TOIS **20**(4), 422–446 (2002)
9. Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., Zhao, L.: Cross-lingual information retrieval with bert. arXiv preprint arXiv:2004.13005 (2020)
10. Karmaker Santu, S.K., Sondhi, P., Zhai, C.: On application of learning to rank for e-commerce search. In: SIGIR, pp. 475–484 (2017)
11. Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019)
12. Li, H., Xu, J.: Semantic matching in search. Found. Trends Inf. Retr. **7**(5), 343–469 (2014)
13. Monz, C., Dorr, B.J.: Iterative translation disambiguation for cross-language information retrieval. In: SIGIR, pp. 520–527 (2005)
14. Nie, J.Y.: Cross-language information retrieval. Synth. Lect. Hum. Lang. Technol. **3**(1), 1–125 (2010)
15. Qin, T., Liu, T.Y., Xu, J., Li, H.: Letor: a benchmark collection for research on learning to rank for information retrieval. Inf. Retrieval **13**(4), 346–374 (2010)
16. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Now Publishers Inc., Delft (2009)
17. Sarvi, F., Voskarides, N., Mooiman, L., Schelter, S., de Rijke, M.: A comparison of supervised learning to match methods for product search. arXiv preprint arXiv:2007.10296 (2020)
18. Sasaki, S., Sun, S., Schamoni, S., Duh, K., Inui, K.: Cross-lingual learning-to-rank with shared representations. In: NAACL, pp. 458–463 (2018)
19. Schamoni, S., Hieber, F., Sokolov, A., Riezler, S.: Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In: ACL, pp. 488–494 (2014)
20. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. NeurIPS **16**, 41–48 (2004)
21. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: Learning semantic representations using convolutional neural networks for web search. In: WWW, pp. 373–374 (2014)

22. Sun, S., Duh, K.: Clirmatrix: a massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In: EMNLP, pp. 4160–4170 (2020)
23. Van Gysel, C., de Rijke, M., Kanoulas, E.: Learning latent vector spaces for product search. In: CIKM, pp. 165–174 (2016)
24. Vaswani, A., et al.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
25. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. JMLR **10**(2), 1 (2009)
26. Yang, Y., et al.: Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. arXiv preprint arXiv:1902.08564 (2019)
27. Zhang, H., et al.: Towards personalized and semantic retrieval: an end-to-end solution for e-commerce search via embedding learning. In: SIGIR, pp. 2407–2416 (2020)
28. Zhang, Y., Wang, D., Zhang, Y.: Neural IR meets graph embedding: a ranking model for product search. In: WWW, pp. 2390–2400 (2019)
29. Zhou, D., Truran, M., Brailsford, T., Wade, V., Ashman, H.: Translation techniques in cross-language information retrieval. CSUR **45**(1), 1–44 (2012)
30. Zhu, H., et al.: Optimized cost per click in Taobao display advertising. In: CIKM, pp. 2191–2200 (2017)