Domain Transfer based Data Augmentation for Neural Query Translation

Liang Yao **Baosong Yang** **Haibo Zhang** **Boxing Chen** **Weihua Luo**
Alibaba Group
Hangzhou, China
{yaoliang.yl, zhanhui.zhb, weihua.luowh}@alibaba-inc.com
{nlp2ct.baosong, chenboxing}@gmail.com

Abstract

Query translation (QT) serves as a critical factor in successful cross-lingual information retrieval (CLIR). Due to the lack of parallel query samples, neural-based QT models are usually optimized with synthetic data which are derived from large-scale monolingual queries. Nevertheless, such kind of pseudo corpus is mostly produced by a general-domain translation model, making it be insufficient to guide the learning of QT model. In this paper, we extend the data augmentation with a domain transfer procedure, thus to revise synthetic candidates to search-aware examples. Specifically, the domain transfer model is built upon advanced Transformer, in which layer coordination and mixed attention are exploited to speed up the refining process and leverage parameters from a pre-trained cross-lingual language model. In order to examine the effectiveness of the proposed method, we collected French-to-English and Spanish-to-English QT test sets, each of which consists of 10,000 parallel query pairs with careful manual-checking. Qualitative and quantitative analyses reveal that our model significantly outperforms strong baselines and the related domain transfer methods on both translation quality and retrieval accuracy.[1]

1 Introduction

Cross-lingual information retrieval (CLIR) can have separate query translation (QT), information retrieval (IR), as well as machine-learned ranking stages. Among them, QT stage takes a multilingual user query as input and returns the translation candidates in language of documents for the downstream retrieval. To this end, QT plays a key role and its output significantly affects the retrieval results (Wu and He, 2010). Recently, neural machine translation (NMT) has shown their superiority in a variety of translation tasks (Ott et al., 2018; Hassan et al., 2018). Several studies begin to explore the feasibility and improvements of NMT for QT task (Sarwar et al., 2019; Sharma and Mittal, 2019).

However, a well-performed NMT model depends on extensive language resources (Popel and Bojar, 2018; Ott et al., 2018; Wan et al., 2020), while there are few available parallel query corpus for neural QT training, limiting further improvement on translation quality. An alternative way to alleviate this problem is to produce pseudo parallel data from large-scale monolingual queries using a translation model, which refers to data augmentation (Sennrich et al., 2016a; Yao et al., 2020). Nevertheless, as a synthesis data producer, the teacher translation system is trained using out-of-domain texts which consists of long and syntax-compliant text rather than short keywords as in queries, as illustrated in Figure 1. Such kind of discrepancy makes the data producer tend to generate readable texts for human but unaware candidates for the downstream retrieval task (Zhou et al., 2012; Sarwar et al., 2019). The augmented data are therefore insufficient to teach QT model how to explicitly generate an in-domain query, leading to weak correlation between translation and retrieval qualities (Yarmohammadi et al., 2019; Rubino, 2020; Bi et al., 2020).

In this paper, we tackle this problem by proposing a domain transfer based data augmentation method. Specifically, we first build the pseudo parallel corpus by translating large-scale source queries into the target language of retrieval documents using a general NMT engine. The out-of-domain candidates is then be revised to in-domain queries by a domain transfer model, thus eliminating mismatch between the

[1]The code and datasets are available at https://github.com/starryskyyl/DTDA.git

Proceedings of the 28th International Conference on Computational Linguistics, pages 4521–4533
Barcelona, Spain (Online), December 8-13, 2020

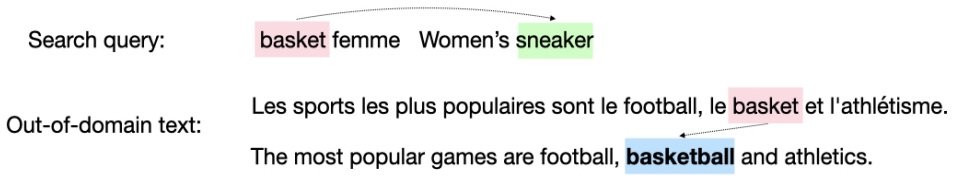

Figure 1: An example to illustrate differences between query and general texts. Considering the text style, the out-of-domain text used to train the teacher translation system is relatively long with complete sentence constituents and syntax, while search query is a short and irregular one. For semantic meaning, the word "basket" is translated into "basketball" in out-of-domain scenario, but "sneakers" in a real-world E-Commerce search engine.

training and inference of QT model. To the best of our knowledge, this is the first study that extend data augmentation with an additional refinement procedure. Moreover, we introduce a novel domain transfer architecture based on a pre-trained cross-lingual language model – Bert (Devlin et al., 2019). Contrast to common refining model that contains two encoders for synthetic sentence pairs and a decoder for generation (Correia and Martins, 2019), our model unifies these components via coordinating and mixing self- and cross-attention layers, resulting in faster processing speed and more adequate architecture to conform to the pre-trained language model. The refinement procedure is fine-tuned by a semi-supervised strategy, which leverages a large amount of monolingual queries and a small number of bilingual query samples that manually generated by human translators.

Researchers may concern about why we employ the forward-translation rather than the backward ones for data augmentation. In a real-world search engine, users with different native languages usually have diverse preferences, resulting in distinct distributions of queries. Deriving data from user queries in target languages leads to potential biases on fields and categories in terms of the search in source language. Besides, noises such as misspelling are common phenomena in queries. Serving these queries as the target training sentence forces QT model to generate worse translations. From this perspective, the domain transfer model to some extent offers the ability on spelling correction. Our extensive analyses can support these hypotheses by showing the superiority of forward data augmentation on QT.

In order to evaluate the effectiveness of the proposed model, we propose French-to-English and Spanish-to-English QT benchmarks, which contains 10,000 manually checked query pairs extracted from real-world E-Commerce website. Experimental results demonstrate that the proposed approach yields better translation and retrieval qualities over the strong Transformer baseline (Vaswani et al., 2017), and related methods that exploited data augmentation (Sennrich et al., 2016a) or domain transfer (Correia and Martins, 2019). In addition, we further conduct experiments on two widely used domain adaptation translation tasks. Results reveal universal-effectiveness of our method by showing consistently improvements on translation performance. Qualitative analyses confirm that the domain transfer model can exactly handle the incorrect and domain-biased cases in pseudo corpus, thus guiding the QT model to learn search-aware translations at the training time.

2 Preliminary

2.1 Neural Machine Translation

NMT uses a single, large neural network to build translation model, aiming to maximize the conditional distribution of sentence pairs using parallel corpus (Sutskever et al., 2017; Bahdanau et al., 2015; Gehring et al., 2017). Formally, the learning objective is to minimize the following loss function over the training corpus $\mathcal{D} = \{(\mathbf{x}^n, \mathbf{y}^n)\}_{n=1}^{N}$, with the size being N:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{x}^n,\mathbf{y}^n)\sim\mathcal{D}}[-\log \mathbf{P}(\mathbf{y}^n|\mathbf{x}^n;\theta)], \tag{1}$$

where $\mathbf{x}^n$ and $\mathbf{y}^n$ indicate the source and target sides of the n-th example in training data. θ denotes the trainable parameters of NMT model. As an advanced NMT model, Transformer (Vaswani et al., 2017)

builds an encoder-decoder framework merely using self-attention networks (Lin et al., 2017; Vaswani et al., 2017) and cross-attention networks (Bahdanau et al., 2015; Luong et al., 2015). The encoder is composed of a stack of N identical layers, each of which has two sub-layers. The first sub-layer is a self-attention network, and the second one is a position-wise fully connected feed-forward network. A residual connection (He et al., 2016) is employed around each of two sub-layers, followed by layer normalization (Ba et al., 2016). Formally, the output of the first sub-layer $\mathbf{C}_e^n$ and the second sub-layer $\mathbf{H}_e^n$ are sequentially calculated as:

$$\mathbf{C}_e^n = \textsc{Ln}(\textsc{Att}(\mathbf{H}_e^{n-1}, \mathbf{H}_e^{n-1}) + \mathbf{H}_e^{n-1}), \quad (2)$$

$$\mathbf{H}_e^n = \textsc{Ln}(\textsc{Ffn}(\mathbf{C}_e^n) + \mathbf{C}_e^n), \quad (3)$$

where $\textsc{Att}(\cdot)$, $\textsc{Ln}(\cdot)$, and $\textsc{Ffn}(\cdot)$ are respectively self-attention mechanism, layer normalization, and feed-forward network with ReLU activation in between.

The decoder is also composed of a stack of N identical layers. In addition to two sub-layers in each decoder layer, the decoder inserts a third sub-layer $\mathbf{D}_d^n$ between self-attention and feed-forward layers to perform cross-attention over the output of the encoder $\mathbf{H}_e^N$:

$$\mathbf{D}_d^n = \textsc{Ln}(\textsc{Att}(\mathbf{C}_d^n, \mathbf{H}_e^N) + \mathbf{C}_d^n). \quad (4)$$

Here, $\textsc{Att}(\mathbf{C}_d^n, \mathbf{H}_e^N)$ denotes attending the top encoder layer $\mathbf{H}_e^N$ with $\mathbf{C}_d^n$ as the output of self-attention layer in decoder. The top layer of the decoder $\mathbf{H}_d^N$ is used to generate the final output sequence.

2.2 Related Work

Query Translation Query translation has attracted increasing attention since its translation quality significantly affects the retrieval results (Wu and He, 2010). Existing studies mainly focus on traditional translation models, e.g. bilingual dictionaries or statistical machine translation systems (Koehn, 2009; Och and Ney, 2002; Gao et al., 2001). Among them, Clinchant and Renders (2007) adapt the initial dictionary to a query-specific dictionary with pseudo relevance feedback methods. Nikoulina et al. (2012) propose to finetune the general translation model using a set of parallel queries. As NMT has shown superiorities in a variety of translation tasks, Sarwar et al. (2019) propose a multi-task learning approach to train a neural-based query translation model with a relevance-based auxiliary task. However, both of these approaches require either expensive language resources or a large amount of parallel data, which are generally inconsistent with the domain of the queries. The lack of parallel query corpus restricts further improvement on translation quality.

Data Augmentation Such kind of low-resource translation task has became an open problem in NMT community. Several data augmentation methods which generate the synthetic parallel corpus $\mathcal{D}_O$ are introduced to alleviate this problem:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{x}^n, \mathbf{y}^n) \sim (\mathcal{D} + \mathcal{D}_O)}[-\log \mathbf{P}(\mathbf{y}^n | \mathbf{x}^n; \theta)]. \quad (5)$$

Sennrich et al. (2016a) propose an effective approach to augment the parallel training data with back-translations of target-side sentences, while Zhang and Zong (2016) assign a self-learning algorithm to make full use of source-side monolingual data and generate the synthetic parallel corpus to enlarge the bilingual training data. Park et al. (2017) use synthetic corpus generated from both sides monolingual sentences as an efficient alternative to real parallel data. Nevertheless, both the translation models used to generate synthesis data are trained utilizing out-of-domain corpus which is composed of long and fluent texts other than short keywords as in queries. This discrepancy makes the synthetic parallel data readable for human but unsuitable for the downstream retrieval task (Zhou et al., 2012; Sarwar et al., 2019). To this end, we attempt to alleviate this problem by exploiting a domain transfer model to refine out-of-domain data to in-domain queries.

Domain Transfer There are several studies focus on domain adaption or automatic post-editing (APE) for machine translation. Typically, a widely used method is to train an NMT model on general domain data and subsequently finetune it on the in-domain data (Luong et al., 2015; Britz et al., 2017). However, such kind of approaches are denounced to lead to catastrophic forgetting (Kirkpatrick et al., 2017) and rely on massive training tricks. Considering the APE model, Correia and Martins (2019) recently propose an encoder-to-decoder architecture which is initialized from Bert (Devlin et al., 2019), demonstrating the effectiveness of the pre-trained language model on APE tasks. Despite their success, this series of work directly revises the final translation, which has a potential risk of generating incorrect modifications.

Different to these studies which leveraging training tricks or post-editing, we propose to transfer the synthetic data to in-domain style before the training of final translation system. In this way, the proposed method not only maintains the diversity of samples, avoiding catastrophic forgetting, but also treats the results of APE as training data, weakening the impact of wrong modification cases.

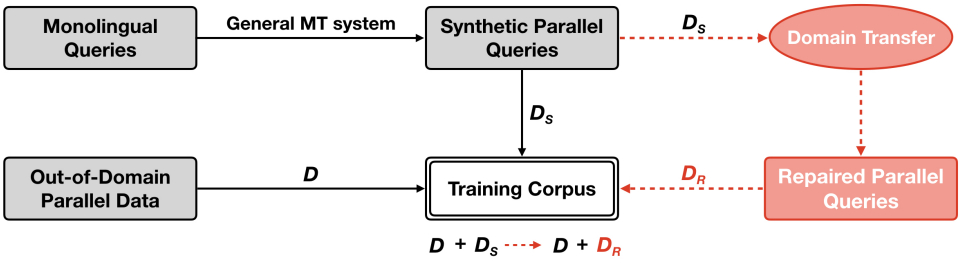

Figure 2: Illustration of the proposed domain transfer based data augmentation method. The black components compose the traditional data augmentation process, while our model (red) revises the synthetic parallel queries $\mathcal{D}_S$ produced by a general translation system with a domain transfer model. The repaired parallel queries $\mathcal{D}_R$ are utilized to augment original out-of-domain parallel data $\mathcal{D}$.

3 Methodology

3.1 Domain Transfer based Data Augmentation

In this section, we described the proposed method. As shown in Figure 2, we extend the traditional data augmentation process with a domain transfer procedure, thus to revise synthetic candidates to search-aware examples. Particularly, the synthetic candidate examples are constructed by translating large-scale source queries into the target language of retrieval documents using a general translation system. The generated candidates are then refined to in-domain queries by a domain transfer model. The repaired parallel queries are utilized to augment original out-of-domain parallel data. Formally, the learning objective of our proposed method is to minimize the following loss function over the merged training corpus built from out-of-domain parallel data $\mathcal{D}$ and the repaired parallel in-domain queries $\mathcal{D}_R$:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{x}^n,\mathbf{y}^n)\sim(\mathcal{D}+\mathcal{D}_R)}[-\log \mathbf{P}(\mathbf{y}^n|\mathbf{x}^n;\theta)]. \quad (6)$$

3.2 Domain Transfer Model

Recent studies have proven the effectiveness of pre-trained language models on various NLP tasks. For example, Devlin et al. (2019) propose M-Bert which is built upon the encoder side of Transformer (Equation 2-3), and trained on large-scale multilingual data. As a representative refining model, Correia and Martins (2019) involve two encoders for synthetic sentence pair and one decoder for generation, both of which are initialized by M-Bert. However, such kind of model assigns additional components for sentence draft encoding which lead to complicated model architecture and reduction of processing speed. Besides, the additional cross-attention layers in decoder (Equation 4) extract features from the top encoding layer, which fail to be initialized from M-Bert. Partially inspired by He et al. (2018) who succeed via sharing layer-wise parameters of encoder and decoder in machine translation tasks, we introduce a novel

domain transfer approach to revise the synthetic parallel queries to in-domain data. As shown in Figure 3, our model unifies components via coordinating and mixing self-attention and cross-attention layers. This offers the model ability to interact features among inputs layer by layer. The proposed architecture conforms to the pre-trained language model, resulting in better initialization from these pre-trained models. In this work, we use M-Bert for initializing the self-attention and mixed-attention parameters. Our architecture contains the following components:

Input Representations Our model accepts three input sources, involving a monolingual query X, its translation result $Y^{'}$ and ground-truth target Y. These sentences are packed together into a single sequence. For a given token, its input representation is constructed by summing the corresponding token, segment, and position embeddings.

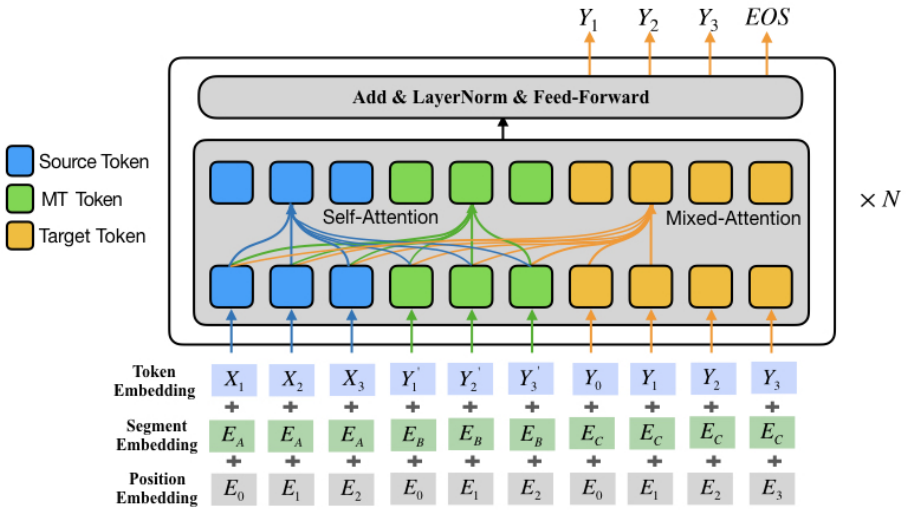

Figure 3: Illustration of the proposed domain transfer model architecture. Our model is composed of a stack of N identical layers. We employ a mixed-attention and layer coordination to share parameters for different inputs. The type of inputs can be identified by segment embeddings.

Self-Attention The representation $\mathbf{H}_X$ of a monolingual query X is paired with the representation $\mathbf{H}_{Y'}$ of its translation result $Y^{'}$. The combined pairs constitutes a new representation $\mathbf{H}_e$ as the input of self-attention network. Similar to traditional Transformer encoder, a position-wise fully connected feed-forward network and residual connection are employed, followed by layer normalization. The attention function is modified as:

$$\mathbf{H}_e^{n-1} = \bigcup\{\mathbf{H}_X^{n-1}, \mathbf{H}_{Y'}^{n-1}\}, \quad (7)$$

$$\mathbf{C}_e^{n} = \textsc{Ln}(\textsc{Att}(\mathbf{H}_e^{n-1}, \mathbf{H}_e^{n-1}) + \mathbf{H}_e^{n-1}), \quad (8)$$

where the union operation $\bigcup$ means combining different types of representations.

Mixed-Attention For each target token Y_j, we merge the self-attention network and cross-attention network of each layer into a unified attention layer called a mixed-attention network. This mixed network breaks the limitation of the scope in each kind of attention networks and allows them to jointly capture features from the representations of the monolingual query X, its translation result $Y^{'}$ as well as previous target tokens $Y_{\leqslant j}$. Contrary to the decoder in Transformer which gets the information from the last layer of the encoder, the proposed model coordinates the layers with same index between source and target. Accordingly, the output of the first sub-layer $\mathbf{C}_{Y_j}^n$ can be expressed as:

$$\mathbf{H}_d^{n-1} = \bigcup\{\mathbf{H}_X^{n-1}, \mathbf{H}_{Y'}^{n-1}, \mathbf{H}_{Y_{\leqslant j}}^{n-1}\}, \quad (9)$$

$$\mathbf{C}_{Y_j}^{n} = \textsc{Ln}(\textsc{Att}(\mathbf{H}_{Y_j}^{n-1}, \mathbf{H}_d^{n-1}) + \mathbf{H}_{Y_j}^{n-1}), \quad (10)$$

where $\mathbf{H}_d^{n-1}$ means combining different types of representations from the monolingual query X, its translation result $Y^{'}$ as well as previous target tokens before position j. Note that, self- and mixed-attention in the same layer share their parameters.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of the proposed domain transfer based data augmentation method (**DTDA**), we conduct experiments on two query translation tasks, including French-English and Spanish-English language pairs. All datasets are tokenized and truecased with the Moses toolkit (Koehn et al., 2007), and splited into sub-word units with a joint BPE model (Sennrich et al., 2016b) with 30K merge operations. The datasets are described as follows:

- **Training Dataset for Domain transfer** In order to train our domain transfer model, a triple dataset need to be constructed, involving a monolingual search query, its machine translation result and ground-truth target. We build this dataset in two ways. Firstly, similar to Negri et al. (2018), we translate the source side of high-quality bilingual query samples (e.g. French-English) created by human translators and utilize the generated translation results to constitute the triple dataset called "Manual data". Secondly, inspired by Junczys-Dowmunt and Grundkiewicz (2016), we create pseudo training triplets by round-trip translations from only monolingual English queries. These monolingual English queries are randomly extracted from the search log of a real-world E-Commerce website (Aliexpress.com[2]). Then, these queries are translated from English to French/Spanish and next backwards from French/Spanish to English. The intermediate French/Spanish translations are preserved. The monolingual English queries are treated as ground-truth target. In this way, we obtain 27.9M artificial training triplets called "Artificial data" for French-English and Spanish-English tasks respectively as shown in Table 1.

- **Training Dataset for Query Translation** In addition to involving the corpus for training domain transfer model, we also extract all available parallel sentences from MultiUN dataset (Eisele and Chen, 2010) as our out-of-domain bilingual corpus. The large-scale monolingual queries are obtained from Aliexpress.com, involving 43.1M french queries and 57.3M spanish queries. These queries are translated by a general machine translation system (Google) to form the original synthetic parallel query pairs. These generated data are then revised by our domain transfer model to in-domain parallel query corpus. The details are presented in Table 1.

- **Development and Evaluation Dataset** We collect 10K French and Spanish search queries from Aliexpress store websites in two countries: Spanish and France, and then manually convert them into English to build parallel query pairs. The data are splited into two equal parts, involving 5K for development and 5K for evaluation. Correspondingly, we translate the source side of above bilingual datasets and then utilize the generated translation results to constitute the triple datasets for our domain transfer model. The detail of development and evaluation data are presented in Table 1. We make them publicly available and contribute to the subsequent researches in the communities of NMT and CLIR.

4.2 Baselines

We compare our model **DTDA** against several effective data augmentation methods as follows:

- **Transformer**: The state-of-the-art NMT model trained on the out-of-domain parallel corpus of MultiUN (Vaswani et al., 2017).

- **Forward-Translation**: The Transformer model augmented with the forward-translation corpus (Zhang and Zong, 2016).

[2]https://www.aliexpress.com/

Task	Training Data					Dev	Test
	Artificial data	Manual data	MultiUN	Synthetic data	Total		
FR-EN	27.9M	0.06M	13.2M	43.1M	84.26M	5,000	5,000
ES-EN	27.9M	0.12M	11.4M	57.3M	96.72M	5,000	5,000

Table 1: Statistics of the datasets for domain transfer(DT) and query translation(QT). M: Million of sentences or queries.

Model	Data		BLEU-4		
	ES-EN	FR-EN	ES-EN	FR-EN	AVG
Transformer	11.4M	13.2M	26.7	40.06	33.38
Forward-Translation (Zhang and Zong, 2016)	96.72M	84.26M	31.32	55.20	43.26
Backward-Translation (Sennrich et al., 2016a)	96.72M	84.26M	30.64	54.48	42.56
Forward & Backward-Translation (Park et al., 2017)	182M	155.3M	30.91	55.31	43.11
Fine-tuning with Forward-Translation Data	96.72M	84.26M	31.83	56.17	44.00
Automatic Post Editing (Correia and Martins, 2019)	96.72M	84.26M	32.92	58.45	45.69
Our Model (DTDA)	96.72M	84.26M	**34.05**†	**60.23**†	47.14

Table 2: BLEU scores on Spanish-English and French-English query translation tasks. Our DTDA model yields better translation performance on the examined tasks than other effective methods. AVG indicates the average BLEU scores on test sets. † represents our system is significantly better than best comparable system ($p<0.01$), tested by bootstrap resampling (Koehn, 2004).

- **Backward-Translation**: The Transformer model reinforced with the backward-translation corpus (Sennrich et al., 2016a).
- **Forward- & Backward-Translation**: The Transformer model trained with both forward and backward translation corpus (Park et al., 2017).
- **Fine-tuning with Forward-Translation Data**: The Transformer model fine-tuned with forward-translation corpus.
- **Automatic Post Editing**: The system utilizing APE to directly revise the final translations produced from the above out-of-domain Transformer model (Correia and Martins, 2019).

4.3 Implementation Details

- **Query Translation**: Neural QT model is based upon the Transformer architecture implemented on the open-source toolkit Tensor2Tensor (Vaswani et al., 2018). Adam optimizer (Kingma and Ba, 2015) is applied with an initial learning rate 0.1. The size of hidden dimension and feed-forward layer are set to 512 and 2048 respectively. Encoder and decoder have 6 layers with 8 heads multi-head attention. Dropout is 0.1 and batch size involves 4096 tokens. Beam size is 4 for inference. We evaluate query translation tasks with tokenized case-insensitive BLEU[3] (Papineni et al., 2002).
- **Domain Transfer**: Instead of a random initialization, the self-attention and mixed attention layers is initialized with the weights of the corresponding self-attention layers of Pre-trained M-Bert (Devlin et al., 2019). Similar to M-Bert, our domain transfer model is composed of 12 layers with hidden size 768 and 12 attention heads. The feed-forward layer size is set to 3072. Adam optimizer is

[3]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

Domain Transfer Model	Languages		Parameters	Tokens/Secs
	ES-EN	FR-EN		
Transformer (Correia and Martins, 2019)	33.62	59.59	251.9M	1.60K
Ours	**34.05**	**60.23**	223.6M	1.95K

Table 3: Comparison of our domain transfer model and APE on two language pairs.

Model	ES-EN			FR-EN		
	MAP	NDCG@10	P@10	MAP	NDCG@10	P@10
Tranformer	32.56	35.10	28.13	20.44	22.64	15.31
Forward-Translation	39.80	42.94	34.56	26.84	29.58	19.99
Our model	**40.01**	**43.02**	**34.73**	**27.00**	**29.78**	**20.18**

Table 4: Comparison of data augmentation methods on two language pairs with respect to retrieval quality. MAP, NDCG@10 and P@10 are common automatic metrics for evaluating retrieval quality.

applied with a learning rate schedule that increases linearly during the first 5,000 steps until 5 * 10-5 and has a linear decay afterwards. Batch size involves 2048 tokens.

4.4 Results

Table 2 summarizes the BLEU scores of different systems on two tasks. As shown, our model DTDA significantly improves translation quality in terms of BLEU, and obtains the best results that gain 1.45 and 3.88 BLEU points over the systems based upon APE and forward-translation respectively. These results demonstrate the effectiveness of the proposed approach. Overall, our experiments indicate the following three points: 1) the model trained with out-of-domain corpus does not perform well on query translation tasks, and 2) synthetic parallel data generated from in-domain monolingual queries significantly improves the translation quality, and 3) traditional refining methods (e.g. APE) which directly post-edit the final translation results obtain better results than forward-translation or backward-translation method.

5 Analysis

In this section, in order to further analyze our proposed model, we explore the effectiveness of several factors, including domain transfer model, retrieval performance and so on. Moreover, we show qualitative analysis on query translation to better understand the advantage of our model.

5.1 Effectiveness of Our Domain Transfer Model

To further investigate the effect of domain transfer model, we compare our proposed model with the popular refining method APE. Different from traditional APE which directly post-edits the final translation result, we utilize the APE to repair the translation corpus and unify the repaired data with out-of-domain corpus for training. As depicted in Table 3, our domain transfer model achieves better translation results with fewer parameters and faster decoding speed.

5.2 Retrieval Performance

We conduct experiments to evaluate whether our proposed DTDA model can improve the retrieval quality of downstream CLIR task. We randomly collected 1k Spanish and French queries from our search log and translated them into English by different translation models. Then, the generated translation results are utilized to retrieve relevant English titles of items from our E-Commerce website. The final retrieval results are scored and manually checked by five bilingual experts. For each query-title pair are labeled by 3 different levels of relevance: bad, good and excellent. MAP, NDCG@10 and P@10 are used as automatic metrics for evaluating retrieval quality. As shown in Table 4, the experimental results show

English	French
Mobile Phone Cases & Covers	Dresses
T-Shirts	Mobile Phone Cases & Covers
Car Stickers	T-Shirts
Action & Toy Figures	Car Stickers
Dresses	Wall Stickers
Covers & Ornamental Mouldings	Necklace

Table 5: Illustration of top 6 categories of English and French search queries.

Query(Language)	Forward-translation(FR-EN)	Backward-translation(EN-FR)
xiomi lumière maison(FR)	xiaomi house light	/
xiomi house light(EN)	/	xiaomi maison lumière

Table 6: Case of misspellings of English and French search queries in E-Commerce search.

that our model achieves significant improvements over the forward-translation based data augmentation method. The results demonstrate that our proposed model not only improves the translation quality of queries, but also promotes the retrieval quality of final CLIR task.

5.3 Forward Translation vs. Backward Translation

To answer why our model employ forward-translation rather than the backward-translation for data augmentation, we investigate the category distribution of different languages in a real-world E-Commerce search engine. As shown in Table 5, the users with different native languages have diverse preferences and potential biases on different categories. Besides, noises such as misspelling are common phenomena in queries. We use Google Spelling Correction Tool to analyze the search queries, and find that more than 20% of English queries are misspelled. Serving these queries as the target training sentence forces query translation model to generate worse translations. The domain transfer model to some extent offers the ability on spelling correction. As illustrated in Table 6, we provide the case of misspellings in French and English in E-Commerce search, and forward-translation and backward-translation provided by our teacher translation model. Taking the backward-translation as the training example may lead the translation system to learn the spelling mistakes of English users (such as "xiomi"), and the forward translation corpus can alleviate such mistakes. Table 2 also shows that the forward-translation method are superior to backward-translation one.

5.4 Qualitative Analysis

We represent the translated results from baselines and our model to explore how our approach ameliorates translation quality. According to Case 1 in Table 7, the french query "talons compense" is translated to "heels offsets" by Transformer and Forward-Translation system, which fail to understand the correct meaning under the context of information retrieval. Our model leverages the domain transferred in-domain queries effectively and figures out the exact meaning of "compense" . With respect to Case 2, the french query "veja chaussur" is failed to be translated to "doka shoes" by APE and our domain transfer model. Our final DTDA model can correctly translate the word "veja" because our method alleviates the modification errors with the union of the repaired data with out-of-domain corpus for training.

5.5 Universality of The Proposed Method

In addition, we further examine the proposed method on two widely used domain translation tasks, i.e., English-German Law and Subtitles. We follow the common experimental setting in Aharoni and Goldberg (2020). Experimental results are concluded in Table 8. For data augmentation, we extract in-domain monolingual data following Tiedemann (2012). Obviously, the domain transfer based data augmentation

	Case1	**Case2**
Source	talons compense	veja chaussur
Reference	wedge heels	veja shoes
Transformer	heels offsets	veja shoes
Forward-Translation	heels offsets	veja shoes
Automatic Post Editing	wedge heels	doka shoes
Our Model (DTDA)	wedge heels	veja shoes

Table 7: Case study on French-English translation results produced by different methods.

Model	Train Data		BLEU-4		
	Law	Subtitles	Law	Subtitles	AVG
Transformer	0.47M	0.5M	41.34	24.37	32.86
Forward-Translation (Zhang and Zong, 2016)	2.39M	5.5M	42.58	24.95	33.77
Our Model (DTDA)	2.39M	5.5M	**43.24**†	**25.6**†	34.42

Table 8: BLEU scores on English-German Law and Subtitles domain adaptation tasks. Our model yields better translation performance on the examined tasks than other effective methods.

approach surpass the others, demonstrating the universal-effectiveness of the propose method.

6 Conclusion

In this paper, we propose a novel data augmentation method based on domain transfer to improve neural QT system. Our contributions are mainly in:

- In order to address the problem of low-resource and alleviate the domain bias in synthetic data, we propose to revise the general domain pseudo training data into search-aware query pairs with a refinement procedure. To the best of our knowledge, this is the first study that employ domain transfer into data augmentation process;
- We design a novel translation domain transfer model, which adopts layer coordination and mixed-attention mechanism, to speed up the processing and sufficiently exploit the parameters in a pre-trained cross-lingual language model;
- We collect two QT tests and make them publicly available, which may contribute to the subsequent researches in the communities of NMT and CLIR;
- Our approach finally outperforms strong Transformer baseline around 8 BLEU on Spanish-to-English QT task and over 20 BLEU on French-to-English QT task. Moreover, our experimental results demonstrate that the proposed method benefits to not only QT tasks but also other domain adaptation translation tasks.

Several interesting improvements can be studied to further strengthen the quality of query translation. For example, it is interesting to exploit the user behavior in search log to extract high-quality translation candidates (Rubino, 2020; Yao et al., 2020). Another promising direction is to combine with other advanced techniques in NMT context (Li et al., 2020; Yang et al., 2020; Wan et al., 2020; Zhou et al., 2020) to further improve the performance of query translation.

Acknowledgements

This work was supported by National Key R&D Program of China (2018YFB1403202). We thank all the anonymous reviewers for their insightful comments.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015*.

Tianchi Bi, Liang Yao, Baosong Yang, and Haibo Zhang. 2020. Constraint Translation Candidates: A Bridge between Neural Query Translation and Cross-lingual Information Retrieval. In *SIGIR eCom'20*.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*.

Stéphane Clinchant and Jean-Michel Renders. 2007. Query translation through dictionary adaptation. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*.

Gonçalo M. Correia and André F. T. Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*.

Jianfeng Gao, Endong Xun, Ming Zhou, Changning Huang, Jian-Yun Nie, and Jian Zhang. 2001. Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint arXiv:1803.05567*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,ACL 2007*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004*.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Jian Li, Xing Wang, Baosong Yang, Shuming Shi, Michael R. Lyu, and Zhaopeng Tu. 2020. Neuron interaction based representation composition for neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*.

Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *13th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2012*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002*.

Jaehong Park, Jongyoon Song, and Sungroh Yoon. 2017. Building a neural machine translation system using only synthetic parallel data. *arXiv preprint arXiv:1704.00253*.

Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Carl Rubino. 2020. The effect of linguistic parameters in clir performance. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech CLSSTS 2020*.

Sheikh Muhammad Sarwar, Hamed R. Bonab, and James Allan. 2019. A multi-task architecture on relevance-based neural query translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.

Vijay Sharma and Namita Mittal. 2019. Refined stop-words and morphological variants solutions applied to hindi-english cross-lingual information retrieval. *J. Intell. Fuzzy Syst.*, 36(3):2219–2227.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2017. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, NIPS 2014*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NIPS 2017*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, François Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018*.

Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. Self-Paced Learning for Neural Madchine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*.

Dan Wu and Daqing He. 2010. A study of query translation using google machine translation system. In *2010 International Conference on Computational Intelligence and Software Engineering*. IEEE.

Baosong Yang, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2020. Improving Tree-based Neural Machine Translation with Dynamic Lexicalized Dependency Encoding. *Knowledge-Based System*.

Liang Yao, Baosong Yang, haibo Zhang, Weihua Luo, and Boxing Chen. 2020. Exploiting neural query translation into cross lingual information retrieval. In *SIGIR eCom'20*.

Mahsa Yarmohammadi, Xutai Ma, Sorami Hisamoto, Muhammad Rahman, Yiming Wang, Hainan Xu, Daniel Povey, Philipp Koehn, and Kevin Duh. 2019. Robust document representations for cross-lingual information retrieval in low-resource settings. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track, MTSummit 2019*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*.

Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, 45(1):1–44.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*.