Exploiting Neural Query Translation into Cross Lingual Information Retrieval

Liang Yao   Baosong Yang   Haibo Zhang*
Weihua Luo   Boxing Chen
Alibaba Group
Hangzhou, China
{yaoliang.yl,yangbaosong.ybs,zhanhui.zhb,weihua.luowh,boxing.cbx}@alibaba-inc.com

ABSTRACT

As a crucial role in cross-language information retrieval (CLIR), query translation has three main challenges: 1) the adequacy of translation; 2) the lack of in-domain parallel training data; and 3) the requisite of low latency. To this end, existing CLIR systems mainly exploit statistical-based machine translation (SMT) rather than the advanced neural machine translation (NMT), limiting the further improvements on both translation and retrieval quality. In this paper, we investigate how to exploit neural query translation model into CLIR system. Specifically, we propose a novel data augmentation method that extracts query translation pairs according to user clickthrough data, thus to alleviate the problem of domain-adaptation in NMT. Then, we introduce an asynchronous strategy which is able to leverage the advantages of the real-time in SMT and the veracity in NMT. Experimental results reveal that the proposed approach yields better retrieval quality than strong baselines and can be well applied into a real-world CLIR system, i.e. Aliexpress e-Commerce search engine.[1]

KEYWORDS

Cross-Language Information Retrieval (CLIR), Clickthrough Data, Query Translation, Neural Machine Translation

1 INTRODUCTION

Query translation serves as an important step in cross-language information retrieval (CLIR)[9]. Generally, CLIR system first translates user query to the language in terms of its index database, the translation quality therefore significantly affects the retrieval results[42].

Recently, neural machine translation (NMT) have shown their superiorities in a variety of translation tasks and even yielded human-level performance [41]. However, existing CLIR systems still apply traditional translation models, e.g. bilingual dictionaries or statistical machine translation systems (SMT) [17, 24], which to some extent restrict the further development of CLIR context. The inapplicability of NMT in CLIR system is mainly lies in three concerns:

C1: Query is usually the distillation of user intention. Contrast to SMT, NMT lacks a mechanism to guarantee all the source words to be translated, resulting in fluent but inadequate translations [13, 21]. Several researchers may doubt on the accuracy and the completeness of a translation generated by NMT model [32, 35, 44].

C2: Recent studies have proven that a well-performed NMT model depends on extensive language resources [25, 31]. The lack of in-domain query pairs may obstruct its training, thus weakening its performance.

C3: As a component in CLIR, query translation has to be real-time, while the advanced NMT model built with multiple network layers are fail to cope with the requirement on processing speed.

To the best of our knowledge, there is few real-world CLIR system exploits NMT and explores its effectiveness. To this end, we aim at resolving the problems list above, thus to make the advanced techniques in terms of NMT contribute the CLIR community.

In response to the first concern (**C1**), we conduct experiments on a widely used multi-lingual search task–CLEF to empirically assess the retrieval quality with respect to SMT and NMT models (Section 5.2). The experimental results shows that queries translated by NMT model perform better on retrieval tasks than that generated by SMT model. This not only dispels the doubt on the inablity of NMT on query translation, our further qualitative analysis but also demonstrates that NMT model can better handle word disambiguation and rare word translation than its SMT counterpart.

Then, we try to further improve neural query translation by tackling the domain-adaptation problem (**C2**) using a novel data augmentation approach. Users across the world issue multilingual queries to the search engines of a website everyday, which form large-scale cross-lingual clickthrough data. Intuitively, when the translation of a query leads the user to click details and even make purchases on recalled items, we attribute the translation pair possesses high quality and enables the expection of users. With the help of such an automatic and low cost quality estimation approach, our model can acquire a large amount of in-domain query translation pairs derived from user behaviors. These training samples can improve the performance of our translation model, of which outputs are further examined by users, fulfilling a positive cycle.

Finally, in order to overcome the problem of latency (**C3**) and deploy NMT model into the real-world e-commerce system, we design a new architecture that features synchronous call to SMT model and asynchronous call to NMT model. Specifically, we employ a cache to store query translation pairs. If the user query fails to hit translation memories, our system requests the real-time SMT engine to return the translation candidates, and asynchronously updates the cache with the translation of NMT model. In this way, our system can jointly profits from the low latency of SMT and the high quality offered by NMT.

We assess the effectiveness of our approaches upon a real-word e-Commerce search engine–Aliexpress. Experimental results demonstrate that the mined data significantly increases the word coverage of our training data, thus to improve the quality of query translation. Moreover, the new architecture makes its average latency comparable to that of SMT (on the magnitude of 10ms), while raises the retrieval accuracy with the help of NMT.

2 BACKGROUND

2.1 Neural Machine Translation

Neural machine translation (NMT) [2] is a recently proposed approach to machine translation which builds a single neural network that takes a source sentence $x = (x_1, ..., x_{T_x})$ as an input and generates its translation $y = (y_1, ..., y_{T_y})$, where x_t and $y_{t'}$ are source and target symbols. Ever since the integration of attention, NMT systems have seen remarkable improvement on translation quality. Most commonly, an attentional NMT consists of three components: (a) an encoder which computes a representation for each source sequence; (b) a decoder which generates one target symbol at a time, shown in Eq.1 ; (c) the attention mechanism which computes a weighted global context with respect to the source and all the generated target symbols.

$$\log p(y|x) = \sum_{t=1}^{T_y} \log p(y_t|y_{t-1}, x) \tag{1}$$

Given N training sentence pairs $(x^i, y^i) \ldots (x^n, y^n) \ldots (x^N, y^N)$, Maximum Likelihood Estimation (MLE) is usually accepted to optimize the model, and training objective is defined as:

$$L_{MLE} = -\sum_{n=1}^{N} \log p(y^n|x^n) = -\sum_{n=1}^{N} \sum_{t=1}^{T_y} \log p(y_t^n|y_{t-1}^n, x^n) \tag{2}$$

Among all the encoder-decoder models, the recently proposed Transformer [41] architecture achieves the best translation quality so far. In this paper, we introduce the most advanced Transformer model architecture into the query translation, which greatly reduces the ambiguity of translation, and improves the quality of retrieval.

The Transformer architecture relies on a self-attention mechanism [20] to calculate the representation of the source and target side sentences, removing all recurrent or convolutional operations found in the previous methods. Each token is directly connected to any other token in the same sentence via self-attention. The hidden state in the Transformer encoder is calculated based on all hidden states of the previous layer. The hidden state h_t^i in a self-attention network is calculated as in Eq.3.

$$h_t^i = h_{t-1}^i + F(self-attention(h_{t-1}^i)) \tag{3}$$

where F represents a feed-forward network with layer normalization and ReLU as the activation function. The decoder additionally has a multi-head attention over the encoder hidden states. For more details, refer to Vaswani [41].
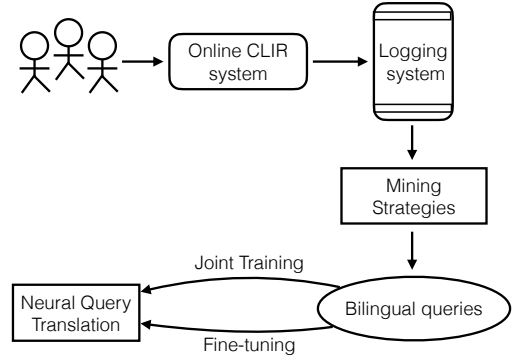

Figure 1: The framework of optimizing neural query translation for CLIR using user clickthrough data.

3 DATA AUGMENTATION WITH CLICKTHROUGH RECORD

In this section, we first introduce the overview of our proposed method. Then, we describe how to record clickthrough data and how to mine in-domain query translation pairs. At last, the method to exploit neural query translation with these mined data will be introduced.

3.1 Overview

Often, the query fed to any information retrieval system can be rather "unnatural", which is very short or in wrong word order or repetitive in meaning [22]. It is difficult for general-purpose machine translation systems which are usually trained on a corpus of standard parallel sentences to deal with such query phrases properly. This problem can be seen as a problem of domain adaptation, where the domain is query [33]. To our knowledge, no suitable corpus of parallel queries is available to train an adapted machine translation system. Although small corpora of parallel queries can be obtained (eg. CLEF tracks) or manually created, it is insufficient for a commercial query translation system.

Intuitively, millions of users across the world issue queries to the search engines of Aliexpress e-commerce website in various languages everyday, which forms large-scale and live cross-lingual clickthrough data. If a large number of users click on one or more items in the search results page of a multilingual query to check the details, or even make a purchase, we believe that the user's expectations have been retained and that the translation of the multilingual query is relatively good. We can utilize the user behavior in the clickthrough data to mine query translation pairs. Such user clickthrough data can be obtained in large quantities and recorded at a very low cost.

Based on this assumption, we propose a novel method to optimize query translation using NMT enhanced with in-domain bilingual translation pairs mined from clickthrough data of cross-lingual search engine. The whole process is presented in Figure 1: a mass of users search, browse, click or purchase on cross-language search websites everyday. These data will be recorded in the logging system. After excavating valuable features from massive user behaviors, we then acquire bilingual query translation pairs through these

features. We then apply these bilingual queries to adapt an NMT system through either data augmentation or fine-tuning method.

3.2 Recording Clickthrough Data

In the scenario of cross-language search on Aliexpress e-commerce website, a complete search process is as follows. After a user inputs a multilingual query, it will first be translated into English. The search engine will then search this English query in the English index database. Then, a search result list in the same language as the original query will be sent to the user who can decide whether to click on any items in the search result or not at all. Such a search process can be formulated as below: where u stands for user, q is the multilingual query, t indicates the translation of the query and c records the click result of user, respectively.

$$d = <u, q, t, c> \qquad (4)$$

We can collect massive data like this in real cross-language search scenario. With the definition of metadata, in the next section we will explain how to mine bilingual queries from massive search logging data.

3.3 Mining Query Translation Pairs

When the user expectation in real cross-language search scenario has been retained, the translation of the multilingual query is to be relatively good. Mining query translation hidden in the user clickthrough data is based on the following two criteria.

The first criterion: When a user clicks on one or a few items in the search result page of a multilingual query to check the details and even makes purchases, we consider the translation of the multilingual query is to be relatively good, and that the user expectation has been preserved. The translation conveys the true expectation of user. When the translation carries critical errors, such as, mistranslation, omission, addition and other translation errors, the search result will be less relevant and the user will click less. Based on this observation, we can define a conversion rate (CTR) of a pair of one multilingual query and its translation $<q, t>$, as a ratio of detail unique view (Duv) to list unique view (Luv). "Detail" is the description page of product. "List" is the list page of search result of a query. $Luv_{<q,t>}$ is defined in Eq.5 as the number of users who have all searched the same multilingual query q. $Duv_{<q,t>}$ in Eq.6 is the sum of users who have also clicked through on at least one items in the search result page.

$$Luv_{<q,t>} = \sum_u \Theta \langle u, q, t \rangle \qquad (5)$$

$$Duv_{<q,t>} = \sum_u \Theta \langle u, q, t \rangle \cdot \Psi(c) \qquad (6)$$

where $\Theta \langle u, q, t \rangle$ in Eq.7 is an indicator that a user u issues a multilingual query q with translation t to the search engine and Ψ in Eq.8 represents whether or not the user clicks the items on the search result page.

$$\Theta \langle u, q, t \rangle = 1 \qquad (7)$$

$$\Psi(c) = \begin{cases} 1, c \geq 1 \\ 0, c = 0 \end{cases} \qquad (8)$$

Once $Luv_{<q,t>}$ and $Duv_{<q,t>}$ are computed, we can work out the conversion rate, i.e., CTR, as Eq.9,

$$CTR_{<q,t>} = \frac{Duv_{<q,t>}}{Luv_{<q,t>}} \qquad (9)$$

Obviously, higher conversion rate $<q, t>$ indicates more reliable translation; otherwise, there might be some problems with the translation or search ranking algorithm. Therefore, translation quality is reliable when the conversion rate is higher than certain threshold value. We may filter out unreliable search query and its translation by following Eq.10 below.

$$CTR_{<q,t>} \geq \eta \qquad (10)$$

The second criterion: Sometimes, click data may be unreliable, especially when a query rarely appears in searches. In this case, whether the user makes clicks or not is extremely random, even when the translation is perfect. We exclude the $<q, t>$ pairs when the original query is rare by forcing $Luv_{<q,t>}$ higher than a threshold.

$$Luv_{<q,t>} \geq \chi \qquad (11)$$

Only those queries $<q, t>$ met two conditions above will be mined and constitute final in-domain bilingual data. Loads of users browse and click everyday on CLIR system, which helps us in choosing and labeling translations of queries. Therefore, we can continuously collect a number of free in-domain bilingual data.

3.4 Training Strategy

In this part, we exploit neural query translation with these mined in-domain translation pairs. Two methods are proposed to enhance the neural translation model, one of which is data augmentation and the other is fine-tuning.

Joint Training (JT):. The mined bilingual queries are mixed with common training data, which are altogether used to train an NMT system from scratch. This is the simplest and most direct approach of applying mined bilingual queries to improve the translation quality of NMT.

Fine-tuning (FT):. Fine-tuning is a fast and simple adaptation method, which has demonstrated substantial improvements in various neural network based models including NMT [5, 11, 36]. To fine tune an NMT, we first train a baseline NMT system with common training data until convergence. Then, we replace the common training data with the mined bilingual queries and continue training until another convergence.

4 ASYNCHRONOUS TRANSLATION ARCHITECTURE

In the previous section, we have deliberately explained the details of our proposed method. Although a large breakthrough has achieved in translation quality, the requirement of low latency in the actual
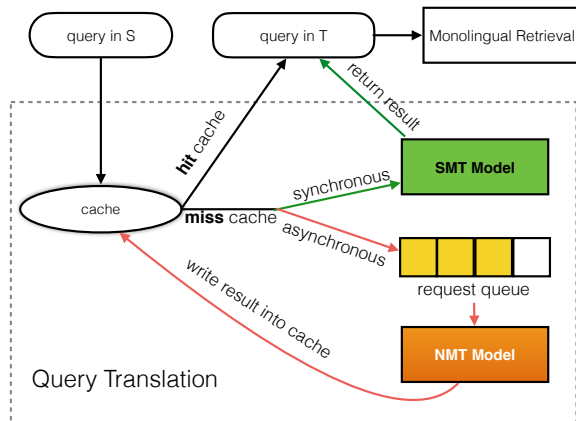
Figure 2: The architecture of the proposed CLIR system. A query in source language (S) from a user will be first translated by the translation module, then retrieve from the commodity library with the well-translated query in target language (T). In the query translation module, we use a cache scheme to bridge asynchronous call and synchronous call. If there is a hit, the model will return the translation result directly. Otherwise, it will synchronously request the SMT engine to acquire the translation result and push the query into a queue being asynchronously translated by NMT.

commercial CLIR scenario prevents the NMT technology from directly applying to the query translation system. It is very difficult for practical application in the commercial CLIR scenario because the translation speed of NMT is about 10 times slower than SMT. Through data analysis on Aliexpress e-commerce website, we found out that there is only a small amount of queries repeatedly assessed by a large number of users every day. The query repetitive rate is up to over 90%.

Given the low latency requirement and high repetitive rate, we propose a new system architecture with the combination of synchronous call to SMT and asynchronous call to NMT for query translation for CLIR. As shown in Figure 2, the architecture is able to maintain the latency at the same millisecond speed of SMT, and complete the user retrieval with a guaranteed quality. The architecture has been broadly tested in the real business scenario of Aliexpress e-commerce website. The process of our architecture is mainly as follows: a query in S from a multilingual user will be first translated by query translation module, then retrieve from the commodity library with the well-translated query in T, and display the retrieved commodity to the user with the language set by the website after document translation. After obtaining the commodity information, the user can browse, click, add to wish-list or purchase the commodity.

In the Query Translation module, we devise a new system architecture that profits on both low latency of SMT and quality offered by NMT. We use a cache scheme to bridge asynchronous call and synchronous call. When a query is requested, it will take priority access to the cache which is only used to store the NMT results,

Table 1: Statistics of the training data we used in CLEF eHealth 2018 Task4.

Language	In-domain	Out-of-domain	Total
CZ-EN	1,145,493	57,279,765	58,425,258
DE-EN	3,036,581	73,518,668	76,555,249
FR-EN	3,418,951	88,213,528	91,632,479

if there is a hit, it will return the translation result directly. Otherwise, it will synchronously request the SMT engine to acquire the translation result at a millisecond level, and return the result directly without writing into the cache. At the same time, it will asynchronously request to add to the queue to be translated by NMT, after translation is completed, the translation result will be stored in the cache. So in the next time when another same query is requested, it can return the NMT result at the millisecond-level. The architecture proposed is capable of meeting over 90% of queries using NMT for the translation result every day in the real business scenario.

5 EXPERIMENTS

In this section, we conducted experiments to answer the following three questions.

1. Is the NMT model better than SMT for query translation to improve retrieval quality in CLIR task?
2. Does the NMT enhanced with mined bilingual in-domain query translation pairs from the user clickthrough data provide improvements in retrieval quality?
3. Can the proposed system architecture provide online services in the real scenario ?

5.1 Setup

5.1.1 Data.

To answer the first question, we undergo the multilingual ad-hoc search task (IRTask4) of CLEF eHealth Consumer Health Search Task 2018 [34, 40] which is similar to the IR tasks in the previous years (2013–2017). In this task, the monolingual English queries were translated by experts into Czech(CZ), French(FR) and German(DE), and the participants are asked to design a search system to retrieve relevant documents to these queries from the English document collection. Two indexes of the document collection in the CLEF 2018 consumer health search task are provided, in the first one, documents are stemmed and a stop-word list is used, while no preprocessing is done in the second index. The collection contains 5,560,074 documents. We use the first index in our experiments. The development set includes queries from the IR task of 2014 (50 queries), 2015 (66 queries), 2016 (300 queries) and 2017 (300 queries), respectively. After removing duplications, there are 416 queries remaining. The test set includes 50 queries from CLEF 2018. The training corpora for NMT contains the UFAL data set and the WMT2018 data set. UFAL Medical Corpus is a collection of parallel corpora aiming at more reliable machine translation of medical texts. UFAL provides in-domain medical data and out-of-domain generic text. WMT2018 provides out-of-domain generic text. The Table 1 shows the composition of the training data.
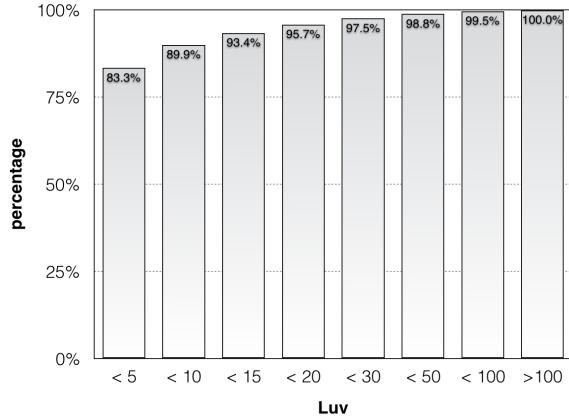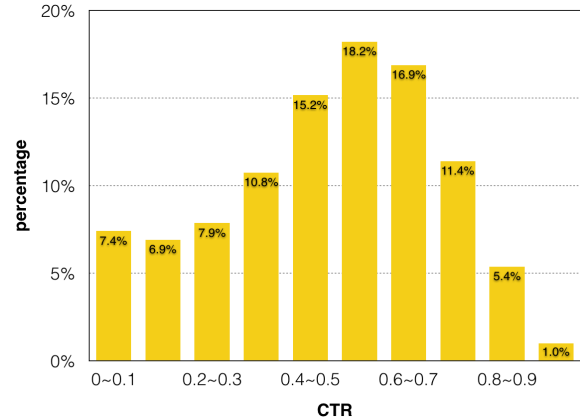
(a) The distribution of queries with different Luv

(b) The distribution of queries with different CTR

Figure 3: (a) The relationship between the ratio of queries and Luv (a) ; and CTR (b). X-axis indicates the threshold of Luv and CTR, respectively. Y-axis presents the ratio of the number of queries to the whole. For instance, the interval of $0 \leq CTR \leq$ reveals that there exists 7.41% queries with CTR being less than 0.1.

Table 2: Statistics of the samples we used for evaluating the proposed data augmentation method. "Top" clickthrough data consists of the queries whose CTR is greater than 0.7, while "Bottom" clickthrough data denotes the group of queries with CTR being less than 0.3.

Type	Source	Size
Training	Top clickthrough data	100k
	Bottom clickthrough data	100k
	WMT 2016 data	2.4M
	E-commerce bilingual data	14M
Validation	E-commerce query with human translation	2k
Test	E-commerce query with human translation	2k

To answer the second question, we collect user clickthrough data from Aliexpress e-commerce website in October 2018 and select data according to the two dimensions of CTR and select data according to Eq.5-11. We use the CLIR system that is serving millions of users through the Aliexpress e-commerce website every day. In terms of Luv, more than 80% of all queries are used by less than 5 people, as summarised in Figure 3a. In consideration for credible statistics, we choose Luv equal or greater than 15 and calculate the CTR distribution of the chosen queries, as shown in Figure 3b. It can be seen that the average value of CTR is greater than 0.5. Setting the CTR threshold at 0.7 and 0.3, we finally selected $100k < q, t >$ pairs as the additional data for training NMT. The main bulk of training data for NMT comes from the WMT2016 Russian-to-English news translation task as well as Russian-to-English e-commerce data. The size of the development set and the test set is 2000, which are extracted from the search log of the Aliexpress e-commerce website in 2017 and produced by human translators. The overall data composition is shown in Table 2.

5.1.2 Systems.

NMT: We adopt the Transformer model with base setting as defined in [41] for experiments. We use BPE[37] to preprocess the source and target texts, forming source-side and target-side dictionaries with 30,000 and 30,000 tokens, respectively. The optimizer used for the MLE training is Adam [16] with initial learning rate at 0.003. We follow the same learning rate schedule as in [41]. During training, roughly 4,096 source tokens and 4,096 target tokens are paired in one mini batch. Each model is trained using 8 NVIDIA Tesla P100 GPUs. For inference, we use beam search with width 6. We run each setting for at least 5 times and report the averaged score on test set. The test set score is chosen via the best configuration based on the validation set.

SMT: We use Moses [18] to train the baseline SMT system. The phrase length is set to 5 and the n-gram of the language model [39] is 3. SMT systems are tuned toward BLEU [28] using MERT [23].

Retrieval System: On the CLEF eHealth 2018 task4, we use the Terrier [26, 27] with Dirichlet model as the retrieval system. On the experiments to verify the impact of the user clickthrough data, we use the online information retrieval system of Aliexpress e-commerce website.

5.2 NMT vs SMT for Query Translation

We performed the experiments for Czech-English, German-English and French-English language pairs of the multilingual ad-hoc search task (IRTask4) of CLEF eHealth Consumer Health Search Task 2018. The following systems are compared to show the impact of NMT .

- Monolingual: retrieval using the manually translated English queries provided with the corpus.
- CUNI: retrieval using 1-best translation of Khresmoi [12] SMT system without re-ranking and query expansion. CUNI [34] is the best system submitted on this task.
- SMT: retrieval using 1-best translation of Moses phrase-based SMT system without re-ranking and query expansion.

Table 3: Experimental Results on CLEF eHealth 2018 Task4 which consists of Czech-English, German-English and French-English cross-lingual search tasks. † represents NMT is significantly better than baseline SMT system.

System	Czech			German			French		
	P@10	MAP	NDCG@10	P@10	MAP	NDCG@10	P@10	MAP	NDCG@10
Monolingual	71.20	27.24	60.46	71.20	27.24	60.46	71.20	27.24	60.46
CUNI	63.00	18.92	51.04	62.00	18.88	53.19	66.80	22.0	56.33
SMT	64.40	21.46	53.28	65.40	21.61	55.85	62.60	20.67	53.05
SMT+BPE	64.00	21.46	53.97	63.20	22.02	53.91	62.00	20.53	52.41
NMT	**66.60**	**22.83**	**55.85**†	**69.00**†	**23.21**	**58.36**	**66.80**†	**22.73**†	**57.10**†

Table 4: Case study on translation examples output by SMT and NMT. "SRC" and "REF" denote the source query and its translation reference, respectively. As seen, NMT translates a Czech phrase "dítě rýma" to "runny nose", while SMT generates "fever", leading to completely irrelevant search results. Considering the German example, NMT is able to translate the word "bewältigungsstrategien" with complex morphology.

Task	Translation Example	P@10	MAP	NDCG@10
Czech⇒English	SRC: čtyřměsíční dítě rýma			
	REF: four month infant running nose	70.00	22.12	54.53
	SMT: four-month-old baby fever	0.00	0.00	0.00
	NMT: a four-month-old runny nose	40.00	3.66	35.56
German⇒English	SRC: angst bewältigungsstrategien			
	REF: anxiety coping skills	40.00	14.88	33.33
	SMT: fear bewältigungsstrategien	0.00	0.00	0.00
	NMT: fear management strategies	30.00	3.58	13.44
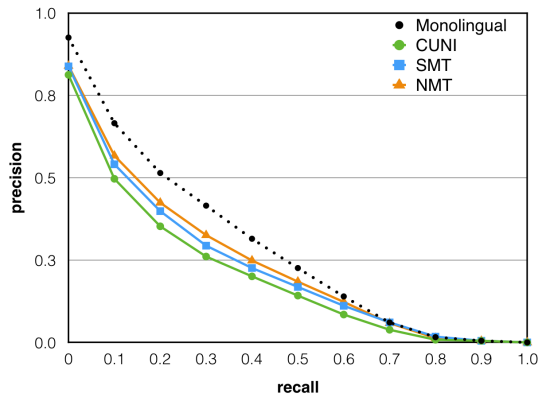
Figure 4: P-R curve of Czech-English in CLEF 2018

- NMT: retrieval using 1-best translation of Transformer system without re-ranking and query expansion.

The experimental results are shown in Table 3. On the P@10, MAP and NDCG@10 indicators, our NMT system has surpassed the baseline SMT and CUNI system. In addition, we show the P-R curve of the Czech-to-English task in Figure 4. The results of German-to-English and French-to-English tasks are similar to that of the Czech-to-English task. For the significance test between SMT and NMT system, we use Wilcoxon test [6] with $\alpha = 0.05$ [14], which is used to compare two independent samples with paired data taken from the same distribution. † represents NMT is significantly better than SMT system. We also try to integrate SMT with BPE. However,

the results show that BPE is less effective when applied to SMT. From these experimental results, we can conclude that NMT can better preserve the query's intention than SMT in query translation.

In order to understand why NMT performs better than SMT in the CLIR task, we analyse the translation results of the test set from Czech-to-English and German-to-English, respectively. As shown in Table 4, the second row is a Czech-English test example. The third row is an example of German-English test. For the Czech phrase "dítě rýma", NMT translates it into "runny nose", while SMT, however, translate this phrase into "fever", leading to completely irrelevant search results. From the point of alleviating translation ambiguity, NMT performs better than SMT even when the query is short and lacks rich context. For the German word "bewälti-gungsstrategien" with complicated morphology, it is difficult to cover the word in the training corpus, so SMT can not translate the word, but NMT uses the sub-word technology to split the word and is able to translate the sub-word units as a sequence. In conclusion, NMT has the natural advantages in disambiguation and translating rare and out-of-vocabulary words over SMT.

Overall, we believe NMT is a better choice for CLIR for the first question. Therefore, we will verify the effect of mined bilingual query translation pairs from the user clickthrough data on the system of NMT.

5.3 Effect of User Clickthrough Data

To answer the second question, we carry out two sets of experiments, where the base training data is of either news articles or e-commerce data and the systems are trained using various combinations of all training data. Transformer is the NMT model and

Table 5: Effectiveness of the proposed data augmentation method on general domain training set. The user clickthrough data is collected from our Aliexpress website. We treat the NMT model trained on WMT 2016 news data as baseline. JT and FT stand for exploiting the augmented data with joint training and fine-tuning, respectively. "coverage" indicates the proportion of words in the test set that appear in the training set. † represents the system is significantly better than baseline.

Train data	System	BLEU-4	P@10	MAP	NDCG@10	Coverage
	Monolingual		71.33	73.77	70.37	
WMT2016	Baseline	27.23	42.66	44.16	42.14	85.22%
WMT2016 + Top clickthrough data	+ JT	34.15	$60.33^{\dagger}$	$61.67^{\dagger}$	$59.84^{\dagger}$	91.70%
WMT2016 + Top clickthrough data	+ FT	**34.40**	$57.00^{\dagger}$	57.30	$56.86^{\dagger}$	91.70%
WMT2016 + Bottom clickthrough data	+ JT	32.57	53.66	56.53	52.61	92.97%
WMT2016 + Bottom clickthrough data	+ FT	31.06	55.33	56.06	55.06	92.97%

Table 6: Effectiveness of the proposed data augmentation method on e-Commerce domain training set. The baseline is trained with our in-house e-commerce data. As seen, the proposed methods can further improve the translation model that trained with in-domain data.

Train data	System	BLEU-4	P@10	MAP	NDCG@10	Coverage
	Monolingual		71.33	73.77	70.37	
E-commerce	Baseline	32.40	59.66	61.04	59.16	91.04%
E-commerce + Top clickthrough data	+ JT	34.78	**66.33**	**67.08**	**66.03**	93.33%
E-commerce + Top clickthrough data	+ FT	**35.48**	64.66	66.40	64.09	93.33%
E-commerce + Bottom clickthrough data	+ JT	33.78	59.66	61.65	59.01	94.96%
E-commerce + Bottom clickthrough data	+ FT	34.54	59.00	60.98	58.28	94.96%

we use the search engine of Aliexpress e-commerce website for document retrieval. The experimental results are shown in Table 5 and 6, where JT stands for joint training and FT for fine-tuning, respectively.

As we can see from the tables, the bilingual query translation pairs mined from the user clickthrough data can significantly improve translation quality and retrieval quality in either general or e-commerce experiments. Such mined bilingual pairs improve the word coverage of the test set in the training set, which is usually beneficial for translation quality. The word coverage indicates the proportion of words in the test set that appear in the training set. Although the word coverage of the test set in the top clickthrough data is lower than that in the bottom clickthrough data, better retrieval quality is observed when the top clickthrough data is used for training the NMT systems. This is mainly because the translations of the queries in the top clickthrough data is more trustworthy and the search engine can find more and better matching results. Nonetheless, the bottom clickthrough data is still somewhat useful in improving retrieval quality. We find that some bottom clickthrough data can also lead to improvement in retrieval quality. The main reason could be either some problems in the ranking process of search engine or that bottom clickthrough data also contains high quality $< q, t >$ pairs.

In both experiment sets, the best system in terms of the quality of query translation is fine-tuned with the top clickthrough data in both sets of experiments, while the best retrieval quality improvement is observed for the NMT system trained with augmented data of the base training data and the top clickthrough data. The observation indicates that the fine-tuning method may overfit the system towards the top clickthrough data, which is actually harmful for

Table 7: Comparison of performance among different systems. "Proportion" presents the proportion of our system calls on SMT and NMT model.

System	Average Latency	Proportion
SMT	9.77ms	100% SMT
NMT	148.43ms	100% NMT
Our architecture	12.87ms	10% SMT + 90% NMT

retrieval quality. This is also in line with previous reports that improving translation quality does not guarantee higher relevance of retrieved documents to the original query in CLIR systems [7, 29].

5.4 Performance of System Architecture

To answer the third question, we collect the user search queries in Russian for one month from Aliexpress e-commerce website and randomly select 30 million queries as test data without deduplication. Since we have mainly innovated in the query translation module among the entire CLIR system, we only test the performance of this module. We compare the following systems SMT, NMT, and our proposed architecture. SMT system is deployed on a five-core CPU server and NMT on an eight-core GPU server of the Tesla P100.

The experimental results are shown in Table 7. The proposed architecture basically achieves the performance of SMT in terms of average latency and has the ability to provide online services. Since the latency of NMT is too high, it is difficult to provide services to users in real business scenarios. Although in recent researches[43, 45], the neural transformer has made remarkable progress in decoding speed, while preserves the translation performance. There are

still great challenges when it is applied to production scenarios that require low-latency and high-throughput decoding. First of all, in the real-world commercial CLIR system, we need to deploy translation services around the world, but we don't have enough GPU machines in all computer rooms. Secondly, if the search query gets longer, the NMT decoding time will not meet the real-time performance requirements of the search.

6 RELATED WORK

A lot of work has been tried to use the user clickthrough data of search engine to optimize a variaty of tasks in CLIR. Hu et al. [4] proposed a method to use a dictionary for finding seed queries to mine URL pair patterns from the click-through data and then used the mined URL pair patterns to generate the candidate query pairs. Ambati et al. [1] used a bilingual dictionary to translate queries in the monolingual click-through data of the target language and built a translation model from the resulting synthesized cross lingual query log to perform CLIR later. Gao et al. [10] mapped the input query of one language to queries of the other language in the query log by such as word translation relations and word co-occurrence statistics for query expansion task. Joachims [15] utilized clickthrough data from the query log of the search engine in connection with the log of links the users clicked to train the ranking algorithm. Kreutzer et al. [19] presented methods to improve quality of title translation from human reinforcement signals which were logged from implicit task-based feedback collected in a cross-lingual search task. In contrast to previous studies, we purposefully mine user clickthrough data to uncover trustable bilingual pairs of multilingual queries and translations produced by a current query translation system of CLIR.

Previous work mainly used bilingual dictionaries or SMT systems to perform query translation. However, most 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. Pirkola et al. [30] were proposed to simplify the process of dictionary-based translation. Gao et al. [8] enhanced the query translation by identifying the phrases using statistical model, then translating the phrases using set of phrase translation patterns and translating the remained words as words. Translating the queries using SMT system has shown potential improvement against other methods in CLIR. Sokolov et al. [38] presented an method to optimise an SMT decoder to immediately output the best translation for CLIR. Bojar et al. [3] tuned SMT systems to produce the best translation in the perspective of humans. Wu et al. [42] conducted a set of CLIR experiments using free available translation system (eg. Google) for translating queries and showed that MT is an excellent tool for the query translation task.

To our knowledge, no such work has been done in studying the effect of NMT on query translation in CLIR, which has proven to be an effective technique in large-scale translation tasks in recent years. Bahdanau et al. [2] proposed attention mechanism into encoder-decoder network and significantly improved the translation quality. The recently proposed Transformer [41] architecture achieves the best translation quality so far which relies on a self-attention mechanism [20] to calculate the representation of the

source and target side sentences, removing all recurrent operations found in the former approach.

7 CONCLUSION

In this paper, we investigate the feasibility of exploiting NMT into CLIR. We first empirically compare the effectiveness of SMT and NMT on query translation. Experimental results dispel the doubt on the inability of NMT on handling short query text. Then, we propose a novel approach to automatically acquiring a large amount of in-domain query translation pairs from the knowledge hidden in the user clickthrough data. The NMT system is able to be enhanced with the mined data using joint training strategy. Besides, to make the NMT technology practical in real-world commercial CLIR system, we devise a new system architecture that profits on both low latency of SMT and quality offered by NMT. To the best of our knowledge, this is the first work to apply the NMT technology to query translation for a real-world CLIR system.

Our experiments show that NMT significantly improves CLIR performance over SMT and outperforms state-of-the-art CUNI system on CLEF 2018 multilingual task 4. Our proposed method is capable of collecting large-scale high-quality in-domain bilingual corpus from implicit user feedback. The extracted in-domain data are integrated into the normal training process of NMT through joint training and fine tuning. At the same time, the proposed architecture solves the problem of NMT online deployment in commercial system.

Although this work shows promising improvement in practice, there still exists several unsolved problems. First, the bilingual data extracted from the user behavior data contains noises which are harmful to the model training. Next, more quality estimation technologies can be explored to further filter out low-quality data. In addition, serving the user behavior as an indicator seems relatively simple, we will consider other indicators for data mining in future. Thirdly, the asynchronous strategy can be treated as a makeshift solution, more speed optimizations from algorithm and engineering perspectives are necessary.

REFERENCES

[1] Vamshi Ambati and U Rohini. 2006. Using monolingual clickthrough data to build cross-lingual search systems. *New Directions in Multilingual Information Access* 28 (2006).

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*. 12–58.

[4] Weizhu Chen and Zheng Chen. 2013. Mining translations of web queries from web click-through data. US Patent 8,543,580.

[5] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *arXiv preprint arXiv:1701.03214* (2017).

[6] Evelyn Fix and JL Hodges Jr. 1955. Significance probabilities of the Wilcoxon test. *The Annals of Mathematical Statistics* (1955), 301–312.

[7] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2009. Evaluating effects of machine translation accuracy on cross-lingual patent retrieval. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (2009), 674–675. https://doi.org/10.1145/1571941.1572072

[8] Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, and Changning Huang. 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.* ACM, 96–104.

[9] Jianfeng Gao, Jian-Yun Nie, and Ming Zhou. 2006. Statistical query translation models for cross-language information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)* 5, 4 (2006), 323–359.

[10] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. 2007. Cross-lingual query suggestion using query logs of different languages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 463–470.

[11] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535* (2015).

[12] Allan Hanbury, Célia Boyer, Manfred Gschwandtner, and Henning Müller. 2011. KHRESMOI: towards a multi-lingual search and access system for biomedical information. *Med-e-Tel, Luxembourg* 2011 (2011), 412–416.

[13] Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with SMT features. In *Thirtieth AAAI conference on artificial intelligence.*

[14] David Hull. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 329–338.

[15] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 133–142.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Philipp Koehn. 2009. *Statistical machine translation.* Cambridge University Press.

[18] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions.* Association for Computational Linguistics, 177–180.

[19] Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can Neural Machine Translation be Improved with User Feedback? *arXiv preprint arXiv:1804.05958* (2018).

[20] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[21] Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track.* 233–243.

[22] 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 *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, 109–119.

[23] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1.* Association for Computational Linguistics, 160–167.

[24] Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics.* Association for Computational Linguistics, 295–302.

[25] 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.*

[26] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. 2005. Terrier information retrieval platform. In *European Conference on Information Retrieval.* Springer, 517–519.

[27] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Christina Lioma. 2006. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of the OSIR Workshop.* 18–25.

[28] 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 on association for computational linguistics.* Association for Computational Linguistics, 311–318.

[29] Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth JF Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, et al. 2014. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial intelligence in medicine* 61, 3 (2014), 165–185.

[30] Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information retrieval* 4, 3-4 (2001), 209–230.

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

[32] 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 (CLSSTS2020).* 1–6.

[33] Shadi Saleh and Pavel Pecina. 2016. Adapting SMT Query Translation Reranker to New Languages in Cross-Lingual Information Retrieval. *In Medical Information Retrieval (MedIR) Workshop, Association for Computational Linguistics* (2016), 1–4.

[34] Shadi Saleh and Pavel Pecina. 2018. Cuni team: Clef ehealth consumer health search task 2018. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum.*

[35] Sheikh Muhammad Sarwar, Hamed Bonab, and James Allan. 2019. A Multi-Task Architecture on Relevance-based Neural Query Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 6339–6344.

[36] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709* (2015).

[37] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* (2015).

[38] Artem Sokolov, Felix Hieber, and Stefan Riezler. 2014. Learning to translate queries for CLIR. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.* ACM, 1179–1182.

[39] Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing.*

[40] Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, et al. 2018. Overview of the CLEF ehealth evaluation lab 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages.* Springer, 286–301.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems.* 5998–6008.

[42] Dan Wu and Daqing He. 2010. A study of query translation using google machine translation system. In *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on.* IEEE, 1–4.

[43] Tong Xiao, Yinqiao Li, Jingbo Zhu, Zhengtao Yu, and Tongran Liu. 2019. Sharing attention weights for fast transformer. *arXiv preprint arXiv:1906.11024* (2019).

[44] 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.* 12–20.

[45] Biao Zhang, Deyi Xiong, and Jinsong Su. 2018. Accelerating neural transformer via an average attention network. *arXiv preprint arXiv:1805.00631* (2018).