Constraint Translation Candidates: A Bridge between Neural Query Translation and Cross-lingual Information Retrieval

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

ABSTRACT

Query translation (QT) is a key component in cross-lingual information retrieval system (CLIR). With the help of deep learning, neural machine translation (NMT) has shown promising results on various tasks. However, NMT is generally trained with large-scale out-of-domain data rather than in-domain query translation pairs. Besides, the translation model lacks a mechanism at the inference time to guarantee the generated words to match the search index. The two shortages of QT result in readable texts for human but inadequate candidates for the downstream retrieval task. In this paper, we propose a novel approach to alleviate these problems by limiting the open target vocabulary search space of QT to a set of important words mined from search index database. The constraint translation candidates are employed at both of training and inference time, thus guiding the translation model to learn and generate well performing target queries. The proposed methods are exploited and examined in a real-word CLIR system–Aliexpress e-Commerce search engine.[1] Experimental results demonstrate that our approach yields better performance on both translation quality and retrieval accuracy than the strong NMT baseline.

KEYWORDS

Query Translation, Cross-lingual Information Retrieval, Neural Machine Translation, Constraint Vocabulary

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 search index for the downstream retrieval. To this end, QT plays a key role and its output significantly affects the retrieval results[23]. In order to improve the translation quality, many efforts have been made based on techniques in machine translation community, e.g. bilingual dictionaries and statistical machine translation [6, 9]. Recently, neural machine translation (NMT) has shown their superiority in a variety of translation tasks [5, 10]. Several studies begin to explore the feasibility and improvements of NMT for QT task [16, 19].

Nevertheless, taking the translation quality as the primary optimization objective for neural query translation may fail to further improve the retrieval performance. Recent studies have pointed out that there seems no strong correlation between translation and retrieval qualities [13, 26]. For example, Fuji et al., [4] empirically investigated this problem, and found the system with the highest human evaluation score in terms of translation, gained the relatively worse retrieval quality. Yarmohammadi et al., [26] also noticed that NMT even has much higher missed detection rate compared to its SMT counterpart, despite its high translation accuracy.

We attribute the mismatch between NMT and CLIR to two reasons. Firstly, a well-performed NMT model depends on extensive language resources [10, 12], while the lack of in-domain query pairs leads existing neural query translation models to be trained using general domain data. This makes a well-trained NMT model fail since the vocabulary and style mismatch between the translated query and terms in search index. On the other hand, the translation model lacks a mechanism to guarantee the produced words to be highly likely in search index at the inference time, resulting in readable texts for human but unaware candidates for the downstream retrieval task [16, 27].

In this paper, we propose to alleviate the mentioned problems by restricting the generated target terms of NMT to constraint candidates of which can be aware by information retrieval system. Since the target search index is built pursuant to the probability distribution of terms in documents, a natural way is to transfer the translation to those target candidates being likely to appear in the retrieval entries. Specifically, given a source query, we mined its constrained target terms according to the distribution of words in the entries clicked by users. The large-scale cross-lingual click-through data on a real-world CLIR engine makes the proposed mining approach feasible and low cost.

We exploit these constraint translation candidates at both of the training and predicting time. For the former, the candidates are served as the smoothed labels during the loss estimation. The NMT model is therefore guided to learn the distribution of search index. For the latter, we limit the output words to the collected candidates with the help of Weighted Softmax. In this way, the search-aware terms offer a bridge between neural query translation and information retrieval.

We build our model upon an advanced neural machine translation architecture–Transformer [3, 21] and evaluate the effectiveness of the proposed approach in a real-word e-Commerce search engine–Aliexpress. Experimental results demonstrate that the proposed method is able to improve the retrieval accuracy, at the same time, maintain the translation quality. The qualitative analysis confirms that our method exactly raises the ability of NMT to generates more suitable target queries for the scenario of e-Commerce search.

*Corresponding author.
[1]We have already exploited NMT into the real-world CLIR system, e.g. Aliexpress. Readers can check their cases on https://aliexpress.com.

2 BACKGROUND

2.1 Neural Machine Translation

Neural machine translation (NMT) [1, 17] 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 [1, 2], 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 $(x^i, y^i) \dots (x^n, y^n) \dots (x^N, y^N)$, Maximum Likelihood Estimation (MLE) is usually accepted to optimize the model, and training objective is defined as:

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

$$= -\sum_{n=1}^{N} \sum_{t=1}^{T_y} \log p(y_t^n|y_{t-1}^n, x^n) \tag{3}$$

Among all the encoder-decoder models, the recently proposed Transformer [21] 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 [8] 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.

$$h_t^i = h_{t-1}^i + F(self{-}attention(h_{t-1}^i)) \tag{4}$$

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 [21].

3 CONSTRAINT TRANSLATION CANDIDATES

In this section, we introduce our proposed method. The neural query translation and information retrieval is bridged with constraint translation candidates. This vocabulary set is mined from parallel corpus and scored according to the term frequency and inverted document frequency in search index. Then, we employ these constraint candidates to guide NMT model to learn and generate the search-aware tokens. Specifically the constrained candidates will be given more weights in training stage. In inference, we will constrain the translation outputs of each query to these candidate vocabularies.

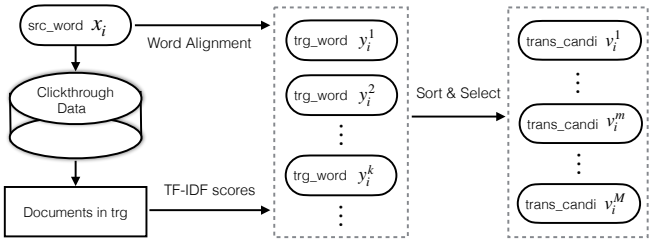

Figure 1: Illustration of the mining method for constraint translation candidates. Our approach first collects the translation candidates using word alignment, which are then sorted and filtered associated with their TF-IDF scores in the set of documents related to the given source word.

3.1 Mining Constraint Candidates

Naturally, an alternative way to select the search-aware translations is to find out those important candidates that likely appear in the retrieval entries, as shown in Figure 1.

Word Alignment

Specifically, given a source word x_i in user query x, we first obtain a set of its possible target tokens C_i^{bi} with its translation possibility distribution in bilingual training corpus. This process can be achieved by a statistical word alignment tool–GIZA++[2] which is able to get alignment distribution between source and target. Generally, GIZA++ implements IBM Models and aligns words based on statistical models. The best alignment of one sentence pair $\widehat{\theta^a}$ is called Viterbi alignment:

$$\widehat{\theta^a} = \underset{\theta^a}{argmax} p_{\widehat{\Psi}}(y^a, \theta^a|x^a) \tag{5}$$

where Ψ can be estimated using maximum likelihood estimation on query translation corpus:

$$\widehat{\Psi} = \underset{\Psi}{argmax} \prod_{s=0}^{S} \sum_{\theta} p_{\Psi}(y^s, \theta|x^s) \tag{6}$$

Here, S is the size of bilingual data. x^s and y^s denotes the source and target sentences, respectively. θ means weights of alignment.

TF-IDF

The candidates C_i^{bi} can be continually scored and filtered according to the distribution of target terms in the entries clicked by users. 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 a recalled item leads the user to click details and even make purchases, we attribute the target tokens in items satisfy the expectation of users. With the help of such an automatic and low cost quality estimation approach, our

[2]https://github.com/moses-smt/giza-pp

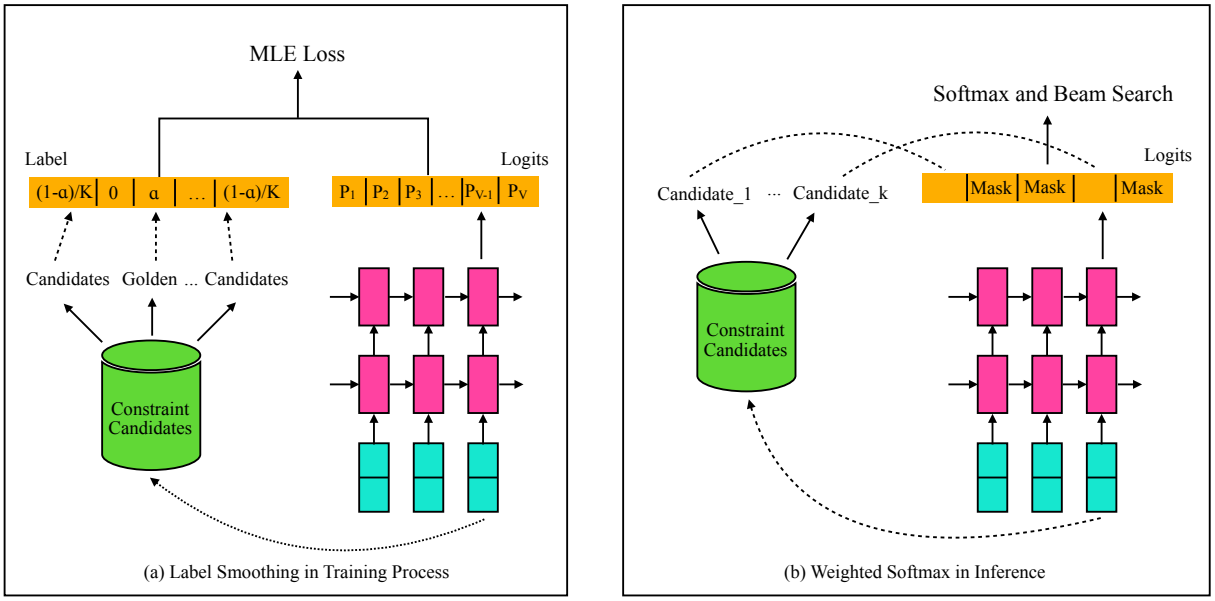

Figure 2: Illustration of training procedure with label smoothing (a) and inference procedure with weighted softmax (b).

model can acquire high quality in-domain translation candidates derived from documents and user behaviors.

From the clickthrough data, we first extract all the documents D_{x_i} that users clicked with any queries contain x_i. Thus, we can use TF-IDF score to identify the importance of each translation candidates in C_i^{bi}:

$$TF - IDF_{y_i^k} = TF_{y_i^k} * IDF_{y_i^k} \tag{7}$$

$$TF_{y_i^k} = \frac{N_{y_i^k}}{\sum_{m=1}^{M} N_{y_i^m}} \tag{8}$$

$$IDF_{y_i^k} = log(\frac{G_Y}{G_{y_i^k} + 1}) \tag{9}$$

where $N_{(*)}$ indicates the frequency that the target term has appeared in D_{x_i}. G_Y denotes the number of documents in D_{x_i} and $G_{y_i^k}$ is the number of documents contain y_i^k.

Different from traditional TF-IDF which calculates scores over all the documents, our approach merely considers the documents that user clicked with a word x_i, thus building correlation among multi-lingual queries and words in documents.

Finally, we can sort the items in C_i^{bi}, and select M words which have the highest scores as constrained translation candidates $V^{candi} = \{v_1, \ldots, v_m, \ldots, v_M\}$. In experiments, we will explore how the size M affects translation quality.

3.2 Training with Label Smoothing

In training process, we use the translation candidates in label smoothing. When calculating the loss of $word_t$, we assign a weight α to the golden label and $1/(1-\alpha)$ to the other constraint translation candidates related to source words equally. With this strategy, we can remove the gap between training and inference. Figure 2 (a) illustrates the training procedure of our proposed method.

In training process, different from traditional MLE, we follow the equations below:

$$L_{MLE_{new}} = \alpha * L_{MLE} + \frac{1}{1-\alpha} * L_{\varphi} \tag{10}$$

$$L_{\varphi} = -\sum_{i=1}^{N}\sum_{t=1}^{T_y}\sum_{m=1}^{M}(logp(v_m|y_{t-1}^n, x^n) \tag{11}$$

where M is the size of words picked from candidates. Contrast to conventional learning objective which merely pays attention to the ground-truth label, we offer the candidates of source words with a loss factor of $1-\alpha$, thus guiding the NMT model to generate the selected words. In our experiments, we empirically set α to 0.6.

3.3 Inference with Weighted Softmax

In NMT, the probability of a prediction is calculated by a non-linear function $softmax$. Given an output hidden state $h \in R^D$ with the hidden size being D, the translation probability of the j-th word in the vocabulary set can be formally expressed as:

$$p(y_j) = \frac{exp(W_j * h + b_j)}{\sum_{k=1}^{|V|} exp(W_k * h + b_k)} \tag{12}$$

where $W \in R^{|V| \times D}$ and $b \in R^V$ are trainable parameter matrix and bias of the vocabulary V, respectively.

As seen, in the conventional approach, all the target words are considered, some of which are completely unrelated to the original query and the downstream search task. Accordingly, an alternative way to assign higher probabilities to constraint translation candidates is to locate factors in $softmax$. In this paper, we apply a more

simple manner that normalizes the probabilities of output words in the proposed constraint space.

$$p(y_j) = \frac{exp(W_j * h + b_j)}{\sum_{k=1}^{|V^{candi}|} exp(W_k * h + b_k)} \quad (13)$$

In this way, the translation model merely calculates the prediction distribution on the constraint translation candidates, thus generating more related tokens for the subsequent task. Figure 2 (b) shows the basic process of translation.

4 EXPERIMENTS

In this section, we conducted experiments on Aliexpress Russian (Ru) to English (En) CLIR engine to evaluate the effectiveness of the proposed method.

4.1 Data

We train our model based on our in-house Ru-En parallel corpus which consists of about 150M general domain sentence pairs. We build the constraint translation candidate by collecting user click-through data from Aliexpress e-commerce website in October 2018. All the Russian and English sentences are tokenized using the scripts in Moses. To avoid the problem of out of vocabulary, the sentences are processed by byte-pair encoding (BPE) [18] with 32K merge operations for all the data. Accordingly, the vocabulary size of Ru and En are set to 30k. 5K queries in search scenarios are randomly extracted and translated by human translators. We treat this dataset as the test set.

4.2 Experimental Setting

We build our model upon advanced Transformer model [21]. Following the common setting, we set the number of layers in encoder and decoder to 6 and hidden size D to 512. We employ multi-head attention with 8 attention heads and 1024 feed-forward dimensions. During training, we set the dropout rate to 0.1. We train our model with parallelization at data batch level with a total batch of 16,384 tokens. For Russia-English task, it takes 300K-400K steps to converge on 4 V100 GPUs. We use Adam optimizer with β_1 = 0.9, β_2 = 0.98 and $\epsilon = 10^{-9}$. We use the same warmup and decay strategy for learning rate as Vaswani et al. [21], with 8000 warmup steps. For evaluation, we use beam search with beam size of 4 and length penalty is 0.6. All the examined models in this paper were re-implemented on the top of our in-house codes based on Tensorflow. We conduct experiments on following models:

- Transformer represents the vanilla NMT baseline with the advanced self-attention-based architecture [21].
- SMT is the phrase-based statistical system of Moses. Our constraint candidates are extracted from the phrase table generated by SMT model.
- +TC denotes the Transformer model enhanced with the proposed methods.

4.3 Translation Quality

In the first series of experiments, we evaluated the impact of different constraint size on the Ru⇒En translation tasks. As shown in Table 1, with the increase of the constraint size, our method consistently improves the translation quality. The result demonstrates that, a small set of constraint translation may miss some of important vocabularies, weakening the generalization ability of the model. The larger constraint size offers a flexible manner to select predictions, thus yields better performance. However, when the size raises to 20, the translation quality reduces. We attribute this to the fact that unrelated candidates makes error propagation from TF-IDF or word alignment, and leads to the decline of translation quality.

Table 1: Ablation study on different constraint size (M in Section 3.1). "Training" and "Inference" denote the constraint size at the training and inference time, respectively. We use BLEU as the assessment metric.

Models	Training	Inference	BLEU (%)
Transformer + TC	5	5	43.81
	10	10	**44.20**
	10	30K	44.06
	30K	10	42.13
	20	20	43.50

Table 2: Main results of the compared models on Ru-En query translation tasks.

Models	BLEU (%)
SMT	38.04
Transformer	43.93
Transformer + TC	**44.20**

Moreover, we also examine the effectiveness of the candidates applied at different stage. As observed, merely constraining the vocabulary size at training time performs better than that at decoding time. We ascribe this to the open problem of exposure bias in deep learning, which is partially caused by the different data distribution between training and decoding. Applying the two strategies jointly yields highest BLEU score, indicating that the two methods are complementary to each other. Finally, we use the best setting, i.e. 10 constraint size for both training and inference, as the default setting in subsequent experiments.

4.4 Main Translation Results

In this section, we evaluate the proposed approach on Ru-En query translation tasks to compare our models with baseline systems, as list in Table 2. Our neural machine translation baseline significantly outperforms the SMT model on such kind of phrase-level text translation task, which makes the evaluation convincing. The results also confirm that the neural query translation model surpasses its SMT counterpart. As seen, the proposed model yields higher BLEU score than the strong baseline system, revealing the effectiveness of our methods to improve the translation quality of query translation.

4.5 Retrieval Performance

We further conduct experiments to learn whether the proposed method can improve the downstream CLIR task. We integrate the

Table 3: Effect of the proposed methods on the downstream information retrieval task.

Metrics	Transformer	Transformer + TC
RECALL	86.69%	87.02%
MAP	29.22	29.47
NDCG@10	35.55	35.71

Table 4: Case study on translation examples output by baseline and the proposed model. "SRC" and "REF" denote the source query and its translation reference, respectively.

SRC	REF	Transformer	Transformer + TC
партфель	briefcase	portfolio	briefcase
поуэрбанк	power bank	poverbank	power bank
мейзу м 6 стекло	meizu m6 glass	maize m 6 glass	meizu m 6 glass

compared query translation models into our CLIR system, and examine the retrieval accuracy of 1612 search queries in 21906 documents. The experimental results are concluded in Table 3. Obviously, on both of RECALL, MAP and NDCG@10 indicators, our model consistently surpass the baseline Transformer model. The results confirm our hypothesis that forcing the query translation model to generate search-ware tokens benefits the retrieval task. The proposed method provides an alternative way to bridge the neural query translation and information retrieval, and offers better recalled items for users.

4.6 Qualitative Analysis

In order to understand how the proposed approach exactly effects the translation and retrieval quality, we analyse the translation results in test set. As shown in Table 4, the case study on Russian to English translation show that, with the help of constraint translation candidates, the quality of translation is indeed improved. For example, in the baseline model which trained with general domain data, the brand of cell phone "meizu" is mistranslated. This is caused by marginal frequency of the token "meizu" in general training data. Thanks to the constraint translation candidates, our model correctly gets the translation. We checked our translation candidate and found that the wrong translation "maize" is not appeared in the list, thus improving the translation quality.

5 RELATED WORK

The correlation between MT system quality and the performance of CLIR system has been studied before. Pecina[11] investigated the effect of adapting MT system to improve CLIR system. They found that the MT systems were significantly improved, but the retrieval quality of CLIR systems did not outperform the baseline system. This means that improving translation quality does not lead to improve the performance of CLIR system. Shadi[14] conducted various experiments to verify that the domain of the collection that CLIR uses for retrieval and the domain of the data that was used to train MT system should be similar as much as possible for better results.

To alleviate the mismatch between translated queries and search index, there are mainly three lines of research works. The first line is re-ranking. Re-ranking takes the alternative translations that are produced by an query translation system, re-ranks them and takes the translation that gives the best performance for CLIR in descending way. Shadi[15] explored a method to make use of multiple translations produced by an MT system, which are reranked using a supervised machine-learning method trained to directly optimize retrieval quality. They showed that the method could significantly improve the retrieval quality compared to a system using single translation provided by MT. The second line is optimizing translation decoder directly. Our work falls into this category. Sokolov[20] proposed an approach to directly optimising an translation decoder to immediately output the best translation for CLIR, which tuned translation model weights towards the retrieval objective and enabled the decoder to score the hypotheses considering the optimal weights for retrieval objective. The last line is multi-task learning which joint multiple tasks into training. Sarwar [16] proposes a multi-task learning approach to train a neural translation model with a Relevance-based Auxiliary Task (RAT) for search query translation. Their work achieves improvement over a strong NMT baseline and gets balanced and precise translations.

6 CONCLUSION

In this paper, we propose a novel approach to tackle the problem of mismatch between neural query translation and cross-lingual information retrieval. We extract a set of constraint translation candidates that contains important words mined from search index database. The constraint translation candidates are incorporated into both of training and inference stages, thus instructing the translation model to learn and generate well performing target queries. Our model is built upon an advanced Transformer architecture and evaluated in a real-word e-Commerce search engine–Aliexpress. Experiments demonstrate that the proposed method can improve the retrieval accuracy and also maintain the translation quality. The qualitative analysis confirms that our method exactly raises the ability of NMT to generates more suitable target queries for the real scenario of e-Commerce search.

As our approach is not limited to information retrieval tasks, it is interesting to validate the similar idea in other cross-lingual tasks that have the mismatch problem. Another promising direction is to design more powerful candidate selection techniques, e.g. calculating the distance between queries using cross-lingual pretrained language models [3]. It is also interesting to combine with other techniques [7, 22, 24, 25] to further improve the performance of neural query translation.

In future, we will continue to focus on how to update the constraint candidate set efficiently and use knowledge of search index to guide query translation through multi-task learning and re-ranking techniques.

REFERENCES

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

[2] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[3] 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*.

[4] 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

[5] 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* (2018).

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

[7] 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 *AAAI*.

[8] 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).

[9] 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.

[10] 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*.

[11] 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.

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

[13] 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.

[14] 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.

[15] Shadi Saleh and Pavel Pecina. 2016. Reranking hypotheses of machine-translated queries for cross-lingual information retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 54–66.

[16] 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.

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

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

[19] Vijay Sharma and Namita Mittal. 2019. Refined stop-words and morphological variants solutions applied to Hindi-English cross-lingual information retrieval. *Journal of Intelligent & Fuzzy Systems* 36, 3 (2019), 2219–2227.

[20] 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.

[21] 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.

[22] Yu Wan, Baosong Yang, Derek F. Wong, Lidia S. Chao, Haihua Du, and Ben CH Ao. 2020. Unsupervised Neural Dialect Translation with Commonality and Diversity Modeling. In *AAAI*.

[23] 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.

[24] Mingzhou Xu, Derek F. Wong, Baosong Yang, Yue Zhang, and Lidia S. Chao. 2019. Leveraging Local and Global Patterns for Self-Attention Networks. In *ACL*.

[25] 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* 188 (2020).

[26] 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.

[27] 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 (2012), 1–44.