Improved Cross-Lingual Question Retrieval for Community Question Answering

Andreas Rücklé
Ubiquitous Knowledge Processing
Lab, Department of Computer Science
Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Krishnkant Swarnkar*
Ubiquitous Knowledge Processing
Lab, Department of Computer Science
Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Iryna Gurevych
Ubiquitous Knowledge Processing
Lab, Department of Computer Science
Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

ABSTRACT

We perform cross-lingual question retrieval in community question answering (cQA), i.e., we retrieve similar questions for queries that are given in another language. The standard approach to cross-lingual information retrieval, which is to automatically translate the query to the target language and continue with a monolingual retrieval model, typically falls short in cQA due to translation errors. This is even more the case for specialized domains such as in technical cQA, which we explore in this work. To remedy, we propose two extensions to this approach that improve cross-lingual question retrieval: (1) we enhance an NMT model with monolingual cQA data to improve the translation quality, and (2) we improve the robustness of a state-of-the-art neural question retrieval model to common translation errors by adding back-translations during training. Our results show that we achieve substantial improvements over the baseline approach and considerably close the gap to a setup where we have access to an external commercial machine translation service (i.e., Google Translate), which is often not the case in many practical scenarios. Our source code and data is publicly available.[1]

CCS CONCEPTS

• **Information systems → Multilingual and cross-lingual retrieval**; *Learning to rank*; *Question answering*.

KEYWORDS

Community Question Answering, Question Retrieval, Cross-lingual Retrieval, Representation Learning, Neural Machine Translation

ACM Reference Format:
Andreas Rücklé, Krishnkant Swarnkar, and Iryna Gurevych. 2019. Improved Cross-Lingual Question Retrieval for Community Question Answering. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308558.3313502

1 INTRODUCTION

Question retrieval is an important task in community question answering (cQA) that allows us to use the existing knowledge which is present in cQA platforms to automatically answer non-factoid questions. We do so by retrieving previously answered questions that are similar to the input question (or query) with the goal of selecting one of their answers as the output [6, 10, 24, 30, 38].

Even though this is a research field with a long-standing history, approaches to question retrieval were traditionally proposed for monolingual scenarios only. However, *cross-lingual* question retrieval is crucial when considering languages other than English because there exists substantially less cQA data in these cases. This is particularly apparent within technical domains, on which we focus in this work, where English platforms such as StackOverflow[2] and AskUbuntu[3] are predominant.

To date, cross-lingual question retrieval has only been explored by two previous works for the language pair English/Arabic and only for general domains, i.e., questions from the Qatar Living forum[4]. However, cross-lingual question retrieval for specialized domains, such as technical cQA, presents us with different challenges because there exists less non-English in-domain data.

First, Joty et al. [22] use adversarial training with feed-forward networks to learn language-invariant feature representations for question pairs. Adversarial training requires in-domain questions from both languages, which are typically not available for specialized domains (e.g., there is no Arabic StackOverflow). And second, Martino et al. [27] apply a mixture of a cross-lingual tree kernel with a dictionary and machine translation with monolingual features. Importantly, the effects from using machine translation were relatively small in their case. However, applying machine translation (MT) to specialized domains is typically more error-prone because large parallel corpora which are used to train MT models mostly contain sentences from general domains [12, 26].

In this work we explore cross-lingual question retrieval for technical domains, i.e., for cQA platforms that cover programming and operating systems topics. Similar to Martino et al. [27] we first use a standard approach to cross-lingual information retrieval, which is to machine translate the query to our target language and then continue with a monolingual question retrieval model [17, 25]. Different to previous work, machine translations would lead to a higher performance decrease for cross-lingual question retrieval in to our specialzed domain. To remedy, we extend this approach with two cross-lingual adaptations that (1) improve the in-domain

*This author is currently at the Indian Institute of Technology (BHU) Varanasi. The present work was performed as part of an internship at UKP Lab.

[1] https://github.com/UKPLab/www19-xling-question-retrieval

[2] www.stackoverflow.com

[3] www.askubuntu.com

[4] www.qatarliving.com/forum

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313502

translation quality and (2) increase the robustness of the question retrieval model to common translation errors.

We achieve this by (a) performing domain-adaptation for the Transformer [42], a state-of-the-art neural machine translation (NMT) approach, by extending the training corpus with synthetic parallel in-domain sentences from cQA platforms. The method we use is known as back-translation in MT literature and has proven to be effective before [3, 11, 23, 31, 37]. In contrast, we test this method extrinsically within question retrieval.

And (b) we increase the robustness of RCNN [24], a state-of-the-art neural question retrieval model, to common translation errors by adding back-translated texts during training. Back-translating is sometimes used to automatically generate paraphrases, e.g., to train paraphrastic sentence embeddings [43] or to obtain more diverse sentences for other (monolingual) tasks [9, 19, 45]. Different to them, we adapt this method to our cross-lingual retrieval scenario.

We perform experiments for the language pair English/German and evaluate our approaches on two datasets from different technical cQA domains—the Askubuntu benchmark [10, 24] and a dataset that we create based on questions from StackOverflow.

Our results show that there exists a substantial cross-lingual performance decrease compared to the monolingual setup, which is in contrast to previous work that performed cross-lingual question retrieval in general domains [27]. In our analysis we observe that this is mostly due to translation errors in our specialized domains. Finally, our proposed adaptations substantially improve the cross-lingual question retrieval performance and close the gap by up to 85% to a scenario where we use a state-of-the-art external commercial MT service (Google Translate) with no adaptations.

Our contributions:

(1) We are the first to explore cross-lingual question retrieval in *technical cQA* and we observe that the retrieval performance is strongly influenced by the translation quality.
(2) We evaluate the impact of NMT domain adaptation extrinsically in cQA question retrieval.
(3) We reduce the cross-lingual performance gap to a setup that has access to an external state-of-the-art commercial MT service with no adaptations by up to 85%. This demonstrates that we can use freely available NMT approaches for cross-lingual question retrieval within technical cQA.

2 RELATED WORK

Early approaches to **monolingual question retrieval** use translation models [2, 20, 44, 49], which can further be enhanced with question category information [6], and topic models [21, 46].

More recent work, especially in the context of the SemEval cQA challenges [30], improve upon this and use tree kernels [8, 34], text alignment features [13], multi-task learning [4], and unsupervised methods with shallow lexical matching and mismatching [47].

In technical cQA domains, neural representation learning methods have proved to be most effective. Dos Santos et al. [10], for example, learn representations of questions with CNNs and compare them with cosine similarity for scoring. Lei et al. [24] propose RCNN, which extends the CNN with a recurrent mechanism (adaptive gated decay). This approach was further extended with question-type information [16], which relies on a question taxonomy for general cQA domains that does not cover the different types of technical questions (e.g., for programming topics).

Other approaches that were applied to technical cQA focus on adversarial domain transfer [38] and on programming-specific association features [48].

Cross-lingual question retrieval has received considerably less attention. Martino et al. [27] explore the language pair English/Arabic within general domains and used a combination of a cross-lingual tree-kernel and machine translation for retrieval. Similarly, Joty et al. [22] explore the language pair English/Arabic within general domains and instead use adversarial training in order to learn language invariant feature representations of question pairs.

Highly related fields are cross-lingual information retrieval and cross-lingual question answering, which was also part of challenges in CLEF 2008 [14], NTCIR-8 [29], and BOLT [39]. Most approaches in this context rely on standard machine translation systems or dictionaries to map questions to their target language [5, 17, 25, 32] and some combine results from multiple MT systems [41], multiple translation outputs from the same MT system [36], or jointly train SMT and IR components [18] to achieve better results.

To the best of our knowledge, we are the first to explore cross-lingual question retrieval in *technical* cQA domains. In contrast to previous work in cross-lingual question retrieval, because specialized domains are more affected by translation errors, we explore the standard approach to cross-lingual information retrieval with two adaptations to (1) improve the translation quality and (2) increase the robustness of a question retrieval model to translation errors.

3 TASK AND METHOD

Given an input question q_0 (the query) and a list of N (potentially) related questions $Q = q_1, q_2, \ldots, q_N$ from a cQA platform that are retrieved with a search engine, the goal is to re-rank this list according to each related question's relevance in regard to q_0.

In our cross-lingual setup, the query is given in a source language L1 (German) and the related questions are in a target language L2 (English). We indicate languages with a superscript, e.g., q_0^{L1} stands for the query in the source language and q_0^{L2} stands for the query in the target language. To perform question retrieval given q_0^{L1} and Q^{L2} we first translate q_0^{L1} to L2 using neural machine translation (NMT). Then we continue with a monolingual L2 question retrieval model to re-rank Q^{L2} according to the translated q_0^{L2}.

This is a standard technique in cross-lingual information retrieval [17, 25] and it has been applied to cross-lingual question retrieval in general domains [27]. We go beyond this by exploring the effects of NMT to the performance of neural question retrieval in specialized domains, i.e., programming and operating systems cQA.

Most importantly, our domains contain a specialized vocabulary which makes NMT—and thus also this standard retrieval approach—more error-prone. This could be even more the case for statistical machine translation [7], which has been used in previous question retrieval work [27]. Thus, in §4 we present two adaptations to improve the translation and ranking performances in our domains.

In the following, we present the NMT model and the neural re-ranking approach. We visualize the full pipeline in Figure 1.

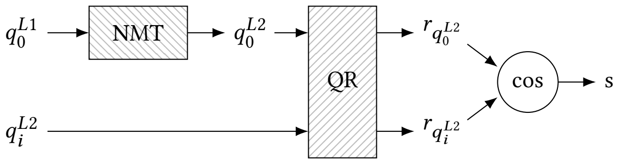

Figure 1: The cross-lingual pipeline. The question retrieval (QR) model learns (monolingual) question representations, which are then compared with cosine similarity to determine a ranking score s. L2 is always English in our setup.

NMT: Synthetic Parallel Data (§4.1)	
orig	Unable to change the launcher icon size
en→de	Nicht in der Lage, die Starter-Symbolgröße zu ändern
QR: Data Augmentation (§4.2)	
orig	How to convert a map to list in java?
en→de	Wie konvertiert man eine Map in Java?
de→en	How to convert a map in Java?

Table 1: Examples for our two data-driven adaptations.

3.1 Neural Machine Translation (NMT)

NMT is a popular research field that has seen rapid progress in recent years. The most effective approaches use the encode-decoder architecture [40] and integrate various forms of attention [1, 15, 42].

In this work, we use Transformer [42], which is a state-of-the-art NMT approach and is based on multi-head self-attention layers. Because it does not make use of recurrent units, Transformer can be efficiently trained on modern hardware while achieving better translation results compared to computationally more expensive models. Edunov et al. [11], for example, use this approach and achieve state-of-the-art results on the WMT'14 en-de benchmark.

For our cross-lingual question retrieval approach we train two models, one for each direction en→de and de→en using the public WMT13 and WMT18 training corpora with more than 4.5M parallel sentences. Even though a small number of sentences in the training data are on technical topics, there is no specific data that is related to technical cQA. To the best of our knowledge there are no publicly available parallel corpora that cover technical questions from cQA.

In addition, for a more thorough comparison, we use an external commercial translation service, namely Google Translate. Even though we expect Google Translate to produce higher quality translations because their model was trained on large amounts of proprietary parallel data, it is not suitable for many practical cQA scenarios due to (1) potential technical restrictions (requiring internet access), (2) legal reasons (confidential information might be transmitted), or (3) associated costs (commercial translation services are not free). The same restrictions do not apply to Transformer models because they are available on-premise.

In the rest of this work we indicate the use of Transformer with TR and the use of Google Translate with GT.

3.2 Question Retrieval (QR)

We use RCNN [24] as a basis for our experiments. This approach combines recurrent units and convolutional neural networks, and it achieves state-of-the-art results on the AskUbuntu question retrieval benchmark [10, 24]. Even though RCNN was recently extended with a separate question type classifier and information from a question taxonomy to improve performances within general domains [16], we use standard RCNN to avoid introducing an additional in-domain question taxonomy (question-type classification would be another source of error in our cross-lingual setup).

For an input question q RCNN learns a fixed-size dense vector representation r_q. This representation is obtained by applying the recurrent CNN sequentially on each token in q where the last output state is considered as the question representation. Model details are given in [24]. This is a strictly monolingual model, which means that we can only learn representations for texts that are in L2.

To rank a related question q_i^{L2} in regard to the query q_0^{L2}, we compare the learned representations with cosine similarity to determine a ranking score s:

$$s = \cos\left(r_{q_i^{L2}}\ ,\ r_{q_0^{L2}}\right)$$

To train RCNN we obtain labeled question paraphrases (also called duplicates) from the cQA platform, i.e., q_0^{L2} and q_+^{L2}. We also obtain one negative sample q_-^{L2} by randomly selecting N questions from the whole corpus and choosing the one with the highest ranking score according to the currently trained model. We then minimize the max-margin hinge loss:

$$\mathcal{L} = \max\left[0,\ m - \cos\left(r_{q_0^{L2}}, r_{q_+^{L2}}\right) + \cos\left(r_{q_0^{L2}}, r_{q_-^{L2}}\right)\right]$$

where m is a non-negative margin.

We additionally perform unsupervised pre-training as proposed in [24], where RCNN is pre-trained in an encoder/decoder setup. The encoder RCNN learns a suitable representation of the question title and/or body and the decoder tries to reconstruct the title form the learned representation.

In general, RCNN can use both the question title and the question body during scoring by applying the element-wise average over the learned representations for both texts. In this work, we use only question titles because it allows us to compare individual sentences for error analysis in our cross-lingual setup. In addition, the performances of models that utilize body information was only marginally better compared to the title-only variant (with pre-training).

4 CROSS-LINGUAL ADAPTATIONS

4.1 In-Domain NMT with Synthetic Parallel Sentences

One disadvantage of training Transformer models on public WMT corpora is that they only contain texts from common domains—they do not include parallel sentences from technical cQA.

Because applying NMT models in out-of-domain scenarios often leads to sub-par translation performances [12, 26], we adapt the Transformer to our specialized domains. There exist several options for domain adaptation, which range from data-centric methods that extend the training corpus to model-centric methods that change the neural network architecture (see [7] for a survey).

In this work we choose a data-centric approach based on back-translation, which has been proven to be effective in both statistical MT [3, 23] and NMT [11, 31, 37] before. Here we train a Transformer model on synthetic parallel sentences which we generate from monolingual in-domain texts. To generate these parallel sentences we apply another Transformer model, which we originally trained on the general WMT corpora, on sentences from technical cQA.

Because we translate queries from L1 to L2 in our cross-lingual question retrieval approach (see Figure 1), we use monolingual L2 data to generate L1 sentences, i.e., the synthetic data is generated by translating in-domain sentences from the target to the source language. This is suitable in our setup because L2 is always English, which means that large-scale in-domain data exists for L2. Also, Lambert et al. [23] show that using translations that were obtained from target to source side results in better domain adaptation compared to using translations from source to target side.

An example of our synthetic data is shown in Table 1. We introduce the source dataset later in §5.1. Based on the synthetic data we then train the in-domain Transformer model, which we denote TR-cQA, by back-translating the generated sentences from L1 to L2. During training we also include the WMT corpora.

4.2 QR Data Augmentation

In addition to adapting the NMT model to our specialized domain, we also enhance the monolingual question retrieval model with the goal of increasing its robustness to common translation errors.

Similar to our NMT adaptation we choose a data-centric approach because it allows us to use the same RCNN network architecture with pre-training as before. In contrast, however, we now augment the training data with back-translated texts, i.e., we translate the titles in the training data from L2 to L1 and back to L2 and use these generated texts during training.

Back-translations are often used as an automatic method to obtain paraphrases, e.g., they can be used to train sentence embeddings [43], to increase the model robustness against adversarial attacks [19], or to obtain more syntactically diverse sentences for training [9, 45]. Here we use them to introduce the QR model to common translation errors. An example is given in Table 1 (bottom), where the aspect of converting sth. *to a list* is lost during translation.

We use GT to obtain back-translations because in initial experiments we found that TR back-translations were too noisy and in many cases they completely altered the meaning of a sentence. The reason is that in the direction en→de we cannot obtain a good TR-cQA model because there exists no large German/L1 technical cQA platform from which we could generate synthetic parallel sentences.

We use back-translated texts in the pre-training and in the regular model training of RCNN (see §3.2). In both phases we add a new training example for every data point by replacing the original question title with its back-translated text.

In the rest of this work we denote the augmented model RCNN-A.

5 EXPERIMENTS

5.1 Experimental Setup

Data. We test our approaches on two datasets from different technical domains. The dataset statistics are shown in Table 2.

Dataset	Number of Examples			N	Paraphrases
	Train	Valid	Test		Valid+Test
AskUbuntu	12,724	189	186	20	6.0
StackOverflow	23,558	1,770	2,779	20	1.3

Table 2: Dataset statistics. N is the number of (potentially) related questions for re-ranking.

First, we use the publicly available *AskUbuntu* benchmark [10] which has been extended with additional manual relevance annotations in [24]. For our cross-lingual setup we got the queries in the validation and test splits translated to German by a German native speaker with proficient English skills and in-domain knowledge.

Second, we obtained questions with labeled duplicates from StackOverflow to create a similar dataset for a different technical domain. We removed all questions that were not tagged with Java or Python so that our dataset has a similar size compared to AskUbuntu. We used BM25 [33] to obtain 20 potentially related questions for every query in our validation and test splits. Questions that were labeled as duplicates by StackOverflow users are considered as relevant (i.e., should be retrieved). In contrast to AskUbuntu we chose much larger validation and test splits to better reduce the effects from random neural network initialization. Due to this larger size, we used Google Translate to obtain the German query questions. Using Google Translate to obtain cross-lingual data has been used before and it was shown that this only minorly impacts the evaluation outcome in other NLP tasks [35].

In the rest of this work we refer to this dataset as *StackOverflow*.

Models and baselines. We evaluate our cross-lingual approach as described in §3 (RCNN TR) with all combinations of our extensions, i.e., RCNN with in-domain NMT (RCNN TR-cQA), with data augmentation (RCNN-A TR) and with both extensions (RCNN-A TR-cQA). We also evaluate the same approach with Google Translate instead of Transformer to provide a comparison to a commercial state-of-the-art MT system (RCNN GT and RCNN-A GT).

Additionally, we evaluate TF*IDF, RCNN, and RCNN-A in the monolingual setup to provide the monolingual baseline and upper-bound.

Training procedure. We train the TR models with the official Transformer implementation[5]. To obtain synthetic in-domain training data for TR-cQA we use TR to translate the titles from the training split of both AskUbuntu and StackOverflow datasets to German (no data from the evaluation splits was used in TR-cQA).

To train RCNN models we use the code provided by Lei et al. [24].

Neural network setup. We use the standard hyperparameters of RCNN except for the batch size and margin parameter *m* where we perform grid search for each model and dataset combination. This is necessary due to the different sizes of the datasets and our data augmentation. The values for the batch size and margin are in {128, 256} and {0.1, 0.5, 1.0} respectively.

For our AskUbuntu experiments we use the 200D word embeddings of Lei et al. [24], which were trained on in-domain data. For StackOverflow we train new 200D word embeddings with Word2Vec [28] on text from stackoverflow.com.

[5] https://github.com/tensorflow/tensor2tensor

Evaluation. For all models we report mean average precision (MAP), mean reciprocal rank (MRR), and precision@1 (P@1; accuracy). To mitigate the effects from random initialization of neural network weights, we report averaged scores over five runs.

5.2 Experimental Results

Table 3 compares the monolingual and cross-lingual performances of the different models for both tasks.

The results show that there exists a considerable performance gap between both setups, e.g., compared to the monolingual RCNN, RCNN TR decreases by 6.6 MAP on StackOverflow and by 3.1 MAP on AskUbuntu.[6] This is in contrast to previous work, where much smaller effects were observed when applying phrase-based MT to cross-lingual question retrieval for general domains [27].

Furthermore, we observe that the translation quality greatly varies in technical cQA and has strong effects on the cross-lingual question retrieval performance—with GT we can achieve substantially better results compared to TR (3.1 MAP on average).

The results also show that our proposed adaptations are effective. They consistently outperform RCNN TR and they substantially close the performance gap to RCNN GT. This is visualized in Figure 2 where we measure the performance improvements in terms of how much they close this performance gap. It shows that adapting TR to technical cQA domains by using synthetic training data has the strongest impact, i.e., it closes the performance gap by over 50% (question retrieval performance increases by 1.8 MAP, 1.9 MRR, and 2.1 P@1, on average). Augmenting the training data of RCNN with back-translated questions further increases the performance and closes the gap by another 19pp to 34pp (+0.6 MAP, +0.9 MRR, and +1.4 P@1). At the same time we observe that data augmentation is also effective for RCNN GT and within the monolingual scenario. The reason for this is that during training the model receives a form of regularization (through noisy samples) and processes additional variations of the monolingual training questions.

The combined model RCNN-A TR-cQA finally substantially reduces the gap to RCNN GT by 75–85%.

6 ANALYSIS

In-domain NMT performance. To test if the improvements with TR-cQA over TR are due to improved translations, we evaluate both models with our human translations from the AskUbuntu evaluation splits and measure their translation performance for de→en. TR-cQA achieves 54.96 BLEU whereas TR has a substantially lower score of 41.70 BLEU. Even though the number of parallel sentences is small (375), the large difference suggests that the improvements of RCNN TR-cQA are indeed due to more accurate translations.

At the same time TR-cQA only decreases marginally on the WMT'14 en/de benchmark (28.04 BLEU vs. 28.30 BLEU).[7]

Error analysis. We compared all cases in which RCNN-A TR-cQA ranks a relevant question first whereas RCNN TR ranks an unrelated question first. Generally, RCNN-A TR-cQA is more robust if the translations slightly differ from the English original query, e.g., when translating "throwing an exception" to "triggering an exception" (which is not semantically equal). In most cases, however, the effects from the improved translations are more apparent.

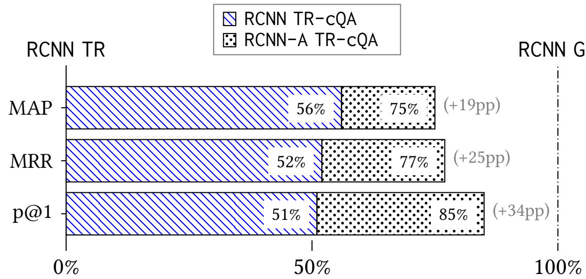

Figure 2: A visualization that shows how much our adaptations close the performance gap between RCNN TR and RCNN GT. Results are averaged over both datasets.

Figure 4 shows a case where TR discards information (i.e., "discrete graphics"). In contrast, TR-cQA generates a translation that is mostly equal to the English source text. During analysis we found that this is a common error of TR, which is greatly improved by TR-cQA due to its better in-domain translation quality. In a small number of cases, however, TR-cQA still discards important information, e.g., when translating long compounds. Another common issue of TR is that it translates words without considering their in-domain meaning, e.g., Figure 5 shows an example where TR translates "null" to "zero".[8] Such errors are often avoided by TR-cQA.

We also compared RCNN GT, RCNN-A TR-cQA, and the monolingual RCNN to understand how we can further improve the cross-lingual question retrieval performances. Most importantly, both RCNN GT and RCNN-A TR-cQA suffer from the same type of problems as described before, but on a smaller scale. This suggests that further improvements are likely with better MT domain adaptation.

7 DISCUSSION

Commercial MT services often achieve better results compared to publicly available models[9] because they have access to large proprietary parallel corpora. In §5.2 we observed a similar trend where RCNN GT still performed better than TR with cross-lingual adaptations. Further, data augmentation in RCNN showed improvements in all cases, including GT. *Why can't we just use Google Translate then?*

There exist several disadvantages when using external MT services in practical scenarios. For example, they can be costly and the data is transferred to third parties. Consider, for example, a non-English company that wants to provide an automatic tool to their software developers that helps them to find related English questions within the StackOverflow data dump (if nothing is found the input question will be sent to colleagues via an internal cQA platform). Because questions of software developers can contain critical and confidential information—e.g., information about the

[6] We also evaluated cross-lingual variants of TF*IDF (with the same NMT approaches), and we tested Spanish RCNN GT with a Spanish version of our StackOverflow dataset. They showed the same trends as the reported RCNN models. We omit the results here for clarity and due to space reasons.

[7] BLEU scores in AskUbuntu are higher because the sentences (i.e., question titles) are considerably shorter compared to WMT'14 sentences.

[8] Null in German typically refers to the number zero. There exist many similar cases, e.g., TR translates (http-) post to "mail".

[9] See, e.g., https://www.deepl.com/quality.html

Setup	**StackOverflow Dev**			**/ Test**			**AskUbuntu Dev**			**/ Test**		
Model	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1	MAP	MRR	P@1
Monolingual												
TF*IDF	58.9	61.0	47.7	54.6	56.7	43.0	53.0	67.5	54.5	52.0	64.5	50.5
RCNN	65.3	67.3	54.5	60.2	62.3	47.9	58.3	72.1	60.1	60.3	73.9	60.6
RCNN-A	65.9	67.9	54.9	60.8	62.9	48.7	58.4	72.8	60.5	60.6	75.7	64.2
Cross-lingual												
RCNN GT	62.3	64.2	50.8	57.7	59.7	45.3	57.3	70.1	56.1	59.0	73.7	60.3
RCNN-A GT	62.9	64.8	51.2	58.5	60.5	46.3	57.8	71.1	57.8	59.2	74.9	62.7
RCNN TR	57.2	59.0	44.5	53.6	55.4	40.4	55.9	68.3	54.3	57.2	70.8	56.9
RCNN TR-cQA	60.2	62.0	48.1	55.4	57.4	42.5	57.0	69.8	55.9	58.3	71.7	58.0
RCNN-A TR	58.1	59.9	45.3	54.4	56.2	41.4	56.1	68.4	54.4	58.3	**73.2**	**60.4**
RCNN-A TR-cQA	**61.2**	**63.1**	**49.3**	**56.0**	**58.0**	**43.1**	**57.6**	**70.5**	**57.4**	**58.4**	72.8	60.3

Table 3: The monolingual and cross-lingual question retrieval performances of the different models.

AskUbuntu

Query: wie kann man diskrete grafik in ubuntu 14.04 deaktivieren *Original text (en):* how to disable discrete graphic in ubuntu 14.04	
RCNN TR-cQA	RCNN TR
Translation: how to disable discrete graphics in ubuntu 14.04	**Translation:** how to disable grafik in ubuntu 14.04
Top-3 retrieved questions	
✓ 14.04 how to disable discrete graphics card	✗ overheating laptop dual ati gpu and discrete
✓ how can i disable ati discrete graphic gpu at startup in ubuntu 14.04 without bios	✓ disabling discrete gpu at startup without system crash
✗ disabling discrete gpu at startup without system crash	✓ how can i disable ati discrete graphic gpu at startup in ubuntu 14.04 without bios

Table 4: A comparison of RCNN TR-cQA and RCNN TR on AskUbuntu, which gives an example of a case where TR discards information during translation.

StackOverflow

Query: try catch exception gibt immer null zurück *Original text (en):* try catch exception always returns null	
RCNN TR-cQA	RCNN TR
Translation: try catch exception always returning null	**Translation:** try catch exception always returns zero
Top-3 retrieved questions	
✓ catching null exception	✗ catch same exception multiple times
✗ exception handling with multiple catch block	✗ exception handling with multiple catch block
✗ catch same exception multiple times	✗ is it expensive to use try - catch blocks even if an exception is never thrown ?

Table 5: A comparison of RCNN TR-cQA and RCNN TR on StackOverflow, which gives an example in-domain vs. out-of-domain translation (see null vs. zero in the translations).

software architecture, stack traces, etc.—it is not suitable to send this data to an external service for translation. In such cases MT needs to be performed with a model that is available on-premise to avoid exposing internal information to third parties.

8 CONCLUSION

In this work we performed cross-lingual question retrieval for *technical cQA* by machine translating the queries to our target language and continuing with a monolingual question retrieval model.

We observed a considerable performance decrease in the cross-lingual scenario compared to the monolingual setup, which is due to a high number of MT errors within our specialized domains.

To remedy this, we first performed domain adaptation for a state-of-the-art NMT model by adding synthetic in-domain parallel sentences to the training corpus. This greatly improved the cross-lingual question retrieval performance due to better in-domain translations. Second, we added back-translated texts during the training of the question retrieval model to increase its robustness, which further improved cross-lingual question retrieval. Finally, the combination of both adaptations substantially closed the performance gap by up to 85% to a model that has access to an external state-of-the-art commercial MT system (with no adaptations), which is often not the case in many practical scenarios.

ACKNOWLEDGEMENTS

This work has been supported by the German Research Foundation (DFG) as part of the QA-EduInf project (grant GU 798/18-1 and grant RI 803/12-1) and by the German Federal Ministry of Education and Research (BMBF) as part of the Software Campus program under the promotional reference 01IS17050.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR 2015)*.

[2] Delphine Bernhard and Iryna Gurevych. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 728–736.

[3] Nicola Bertoldi and Marcello Federico. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 182–189. http://aclweb.org/anthology/W09-0432

[4] Daniele Bonadiman, Antonio Uva, and Alessandro Moschitti. 2017. Effective shared representations with Multitask Learning for Community Question Answering. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Association for Computational Linguistics, 726–732. http://aclweb.org/anthology/E17-2115

[5] Gosse Bouma, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jörg Tiedemann. 2008. Question Answering with Joost at CLEF 2008. In *9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008)*. 257–260.

[6] Xin Cao, Gao Cong, Bin Cui, Christian S Jensen, and Quan Yuan. 2012. Approaches to exploring category information for question retrieval in community question-answer archives. *ACM Transactions on Information Systems (TOIS)* 30, 2 (2012), 7.

[7] Chenhui Chu and Rui Wang. 2018. A Survey of Domain Adaptation for Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Association for Computational Linguistics, 1304–1319. http://aclweb.org/anthology/C18-1111

[8] Giovanni Da San Martino, Alberto Barrón Cedeño, Salvatore Romeo, Antonio Uva, and Alessandro Moschitti. 2016. Learning to re-rank questions in community question answering using advanced features. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM 2016)*. ACM, 1997–2000.

[9] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to Paraphrase for Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. Association for Computational Linguistics, 875–886. https://doi.org/10.18653/v1/D17-1091

[10] Cicero Dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. Learning Hybrid Representations to Retrieve Semantically Equivalent Questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 694–699. https://doi.org/10.3115/v1/P15-2114

[11] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381* (2018).

[12] M. Amin Farajian, Marco Turchi, Matteo Negri, Nicola Bertoldi, and Marcello Federico. 2017. Neural vs. Phrase-Based Machine Translation in a Multi-Domain Scenario. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Association for Computational Linguistics, 280–284. http://aclweb.org/anthology/E17-2045

[13] Simone Filice, Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2016. KeLP at SemEval-2016 Task 3: Learning Semantic Relations between Questions and Answers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, 1116–1123. https://doi.org/10.18653/v1/S16-1172

[14] Pamela Forner, Anselmo Peñas, Eneko Agirre, Iñaki Alegria, Corina Forăscu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, et al. 2008. Overview of the clef 2008 multilingual question answering track. In *9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008)*. 262–295.

[15] Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344* (2016).

[16] Deepak Gupta, Rajkumar Pujari, Asif Ekbal, Pushpak Bhattacharyya, Anutosh Maitra, Tom Jain, and Shubhashis Sengupta. 2018. Can Taxonomy Help? Improving Semantic Question Matching using Question Taxonomy. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Association for Computational Linguistics, 499–513. http://aclweb.org/anthology/C18-1042

[17] Sven Hartrumpf, Ingo Glöckner, and Johannes Leveling. 2008. Efficient question answering with question decomposition and multiple answer streams. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 421–428.

[18] Felix Hieber and Stefan Riezler. 2015. Bag-of-Words Forced Decoding for Cross-Lingual Information Retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2015)*. Association for Computational Linguistics, 1172–1182. https://doi.org/10.3115/v1/N15-1123

[19] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*. Association for Computational Linguistics, 1875–1885. https://doi.org/10.18653/v1/N18-1170

[20] Jiwoon Jeon, W Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM 2005)*. 84–90.

[21] Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. 2012. Question-answer Topic Model for Question Retrieval in Community Question Answering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2471–2474. https://doi.org/10.1145/2396761.2398669

[22] Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language Learning with Adversarial Neural Networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, 226–237. https://doi.org/10.18653/v1/K17-1024

[23] Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on Translation Model Adaptation Using Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 284–293. http://aclweb.org/anthology/W11-2132

[24] Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Katerina Tymoshenko, Alessandro Moschitti, and Lluis Marquez. 2016. Semi-supervised Question Retrieval with Gated Convolutions. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2016)*, 1279–1289. https://doi.org/10.18653/v1/N16-1153

[25] Chuan-Jie Lin and Yu-Min Kuo. 2010. Description of the NTOU Complex QA System.. In *NTCIR-8 Workshop*. 47–54.

[26] Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of IWSLT*. 76–79.

[27] Giovanni Da San Martino, Salvatore Romeo, Alberto Barrón-Cedeño, Shafiq R. Joty, Lluís Màrquez i Villodre, Alessandro Moschitti, and Preslav Nakov. 2017. Cross-Language Question Re-Ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*.

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS 2013)*. 3111–3119.

[29] Teruko Mitamura, Eric Nyberg, Hideki Shima, Tsuneaki Kato, Tatsunori Mori, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin, Tetsuya Sakai, Donghong Ji, et al. 2008. Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR-8)*.

[30] Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 Task 3: Community Question Answering.

[31] Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating Backtranslation in Neural Machine Translation. *arXiv preprint* (2018). http://arxiv.org/abs/1804.06189

[32] Amir Pouran Ben Veyseh. 2016. Cross-Lingual Question Answering Using Common Semantic Space. In *Proceedings of TextGraphs-10: the Workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, 15–19. https://doi.org/10.18653/v1/W16-1403

[33] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 232–241.

[34] Salvatore Romeo, Giovanni Da San Martino, Alberto Barrón-Cedeno, Alessandro Moschitti, Yonatan Belinkov, Wei-Ning Hsu, Yu Zhang, Mitra Mohtarami, and James Glass. 2016. Neural attention for learning to rank questions in community question answering. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*. 1734–1745.

[35] Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated Power Mean Embeddings as Universal Cross-Lingual Sentence Representations. *arXiv* (2018). https://arxiv.org/abs/1803.01400

[36] Shadi Saleh and Pavel Pecina. 2016. Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer International Publishing, 54–66.

[37] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Association for Computational Linguistics, 86–96. https://doi.org/10.18653/v1/P16-1009

[38] Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial Domain Adaptation for Duplicate Question Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Association for Computational Linguistics, 1056–1063. http://aclweb.org/anthology/D18-1131

[39] Ian Soboroff, Kira Griffitt, and Stephanie Strassel. 2016. The BOLT IR Test Collections of Multilingual Passage Retrieval from Discussion Forums. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2016)*. ACM, 713–716.

[40] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS 2014)*. 3104–3112.

[41] Ferhan Ture and Elizabeth Boschee. 2016. Learning to Translate for Multilingual Question Answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*. Association for Computational Linguistics, 573–584. https://doi.org/10.18653/v1/D16-1055

[42] 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 (NIPS 2017)*. 5998–6008.

[43] John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*. Association for Computational Linguistics, 274–285. http://aclweb.org/anthology/D17-1026

[44] Xiaobing Xue, Jiwoon Jeon, and W Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008)*. ACM, 475–482.

[45] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. *International Conference on Learning Representations (ICLR 2018)* (2018).

[46] Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. 2014. Question Retrieval with High Quality Answers in Community Question Answering. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM 2014)* (2014), 371–380.

[47] Minghua Zhang and Yunfang Wu. 2018. An Unsupervised Model with Attention Autoencoders for Question Retrieval. In *Proceedings of the ThirtySecond AAAI Conference on Artificial Intelligence (AAAI 2018)*. 4978–4986.

[48] Wei Emma Zhang, Quan Z Sheng, Jey Han Lau, Ermyas Abebe, and Wenjie Ruan. 2018. Duplicate Detection in Programming Question Answering Communities. *ACM Transactions on Internet Technology (TOIT)* 18, 3 (2018), 37.

[49] Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-Based Translation Model for Question Retrieval in Community Question Answer Archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011)*. Association for Computational Linguistics, 653–662. http://aclweb.org/anthology/P11-1066