Incremental Domain Adaptation for Neural Machine Translation in Low-Resource Settings

Marimuthu Kalimuthu **Michael Barz** **Daniel Sonntag**
German Research Center for Artificial Intelligence (DFKI),
Saarland Informatics Campus D3_2, 66123 Saarbrücken
{marimuthu.kalimuthu,michael.barz,daniel.sonntag}@dfki.de

Abstract

We study the problem of *incremental domain adaptation* of a generic neural machine translation model with limited resources (e.g., budget and time) for human translations or model training. In this paper, we propose a novel query strategy for selecting "unlabeled" samples from a new domain based on *sentence embeddings* for Arabic. We accelerate the fine-tuning process of the generic model to the target domain. Specifically, our approach estimates the informativeness of instances from the target domain by comparing the distance of their sentence embeddings to embeddings from the generic domain. We perform machine translation experiments (Ar-to-En direction) for comparing a *random sampling* baseline with our new approach, similar to active learning, using two small update sets for simulating the work of human translators. For the prescribed setting we can save more than 50% of the annotation costs without loss in quality, demonstrating the effectiveness of our approach.

1 Introduction

Neural Machine Translation (NMT) is the task of translating text from one language (source) to another (target) using, most commonly, Recurrent Neural Networks (RNN), specifically the Encoder-Decoder or Sequence-to-Sequence models (Sutskever et al., 2014; Cho et al., 2014). Recently, NMT has become a quite popular and effective alternative to traditional Phrase-Based Statistical Machine Translation (PBSMT) (Koehn et al., 2003). Major problems that arise include very high cost of training NMT models for new domains and that abundant parallel corpora are required for this task: the standard encoder-decoder models with attention have been shown to perform poorly in low-resource settings (Koehn and Knowles, 2017). Sufficient data might not be available for all languages due to resource restrictions, particularly for resource-poor languages. Hence, we are in need of cost-effective adaptation techniques that transfer existing knowledge to new domains as much as possible.

A recently proposed approach for domain adaptation filters generic corpora based on sentence embeddings of a potentially low amount of in-domain samples to train domain-specific models from scratch (Wang et al., 2017). However, the problem of time- and resource-consuming training still remains which is unsuitable for incremental model updates.

Fine-tuning can accelerate the training process because it transfers knowledge from a pre-trained *generic* model to a new domain and, hence, requires less parallel training samples. However, respective differences in contents and writing style can reduce machine translation quality, if they are not properly addressed.

Recent approaches include fine-tuning with mixed batches containing in- and out-of-domain samples (Chu et al., 2017) and with different regularization methods for differing amounts of new samples for English → German and English → Russian (Barone et al., 2017). The findings of Barone et al. (2017) suggest that there is an "approximately logarithmic relation between the size of in-domain training set and improvement in BLEU score". We want to find out whether incremental model training can be accelerated using an advanced query strategy for sample selection. Previous works on incremental machine translation include cache-based computer aided translation tools (Nepveu et al., 2004), active learning techniques for interactive statistical machine translation (González-Rubio et al., 2012), interactive visualizations for understanding and manipulating attention weights and beam search parameters in NMT (Lee et al., 2017), and domain adap-

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 1–10
Florence, Italy, August 1, 2019.

tation through user interactions (Peris and Casacuberta, 2018).

In this work, we implement a new query strategy for selecting "unlabeled" instances from a target domain and investigate its effect on fine-tuning a generic NMT model. We borrow techniques and terms from the active learning domain (Settles, 2010): a *query strategy* is a method for selecting instances from a pool of unlabeled data that lead to a high information gain when used for training the machine learning model under consideration. Selected instances are labeled by an oracle which can be a human. Iteratively including the most informative instances, labeled on demand, has been shown to increase the model performance while using the same amount of training data. Our proposed methods for domain adaptation in NMT include query strategies that consider untranslated sentences as unlabeled instances. We simulate a human oracle by using parallel corpora in the evaluation, but we do not consider incremental updates for the query strategy. This is of interest for crowd-based domain-adaptation with limited resources as described in (Barz et al., 2018b), in particular, because our method only requires monolingual data for filtering (see Figure 1).

We compare random sampling as a naïve baseline strategy with our novel method based on distances between sentence embeddings. We estimate the informativeness of instances from the target domain by comparing the distances of their sentence embeddings to the embeddings of the generic domain. For computing the sentence embeddings, we present AraSIF: we adapted the methodology presented by Arora et al. (2017), which is known to capture the semantics of sentences well, to work with Arabic. In our experiments, we use existing parallel corpora for simulating human workers: The MEDAR[1] and GlobalVoices dataset (Tiedemann, 2012) are considered as new target domains which mainly concern the domain of *climate change* and *politics*, respectively. The LDC Newswire parallel corpus is used as the dataset for training *generic* domain model. We fine-tune this *generic* NMT model using different amounts of samples from a new domain and varying training epoch settings while observing the BLEU score (Papineni et al., 2002) on a held-out in-domain test set. Our hypothesis is that the proposed novel query strategy can effectively reduce the number of fine-tuning samples required without hampering the translation quality when compared to the baseline.

The remainder of this paper is organized as follows: Section 2 provides an overview on related works, section 3 describes the NMT system and considered query strategies. In section 4, we describe our experiment, and we report the results in section 5. The results are discussed in section 6 and we conclude our work in section 7.

2 Related Work

Almahairi et al. (2016) presented their first result on AR–EN bidirectional NMT, showing that NMT models outperform traditional PBSMT models when they are tested on out-of-domain test data. This result motivates us to study domain adaptation of NMT models rather than PBSMT models.

Several approaches are proposed for domain adaptation in the context of statistical and neural machine translation. Wang et al. (2017) show a way to adapt existing corpora to new domains using learned sentence embeddings for the source language of an NMT model to identify training samples that are close to the new domain. This method allows us to train NMT models for new domains without requiring a parallel corpus in that domain, but models need to be trained from scratch. Chu et al. (2017) present a method called "*mixed fine-tuning*" where fine-tuning is performed with mini-batches composed of a mix of in- and out-of-domain parallel samples to address the problem of overfitting to the new domain. Barone et al. (2017) investigate regularization methods for domain adaptation in NMT. Their findings indicate that BLEU scores increase logarithmically with an increasing amount of in-domain training data. Peris and Casacuberta (2018) implement an online domain adaptation method based on user interactions on the sub-word level. In an experiment, simulating such interactions with available public corpora, they could show that their online learning approach successfully improves word error rates for EN-to-DE and EN-to-FR translations.

González-Rubio et al. (2012) present different active learning techniques that shall reduce human workload in interactive statistical machine translation. They consider three query strategies for selecting the most informative sentences for being translated by humans: a random sampling base-

[1] http://medar.info

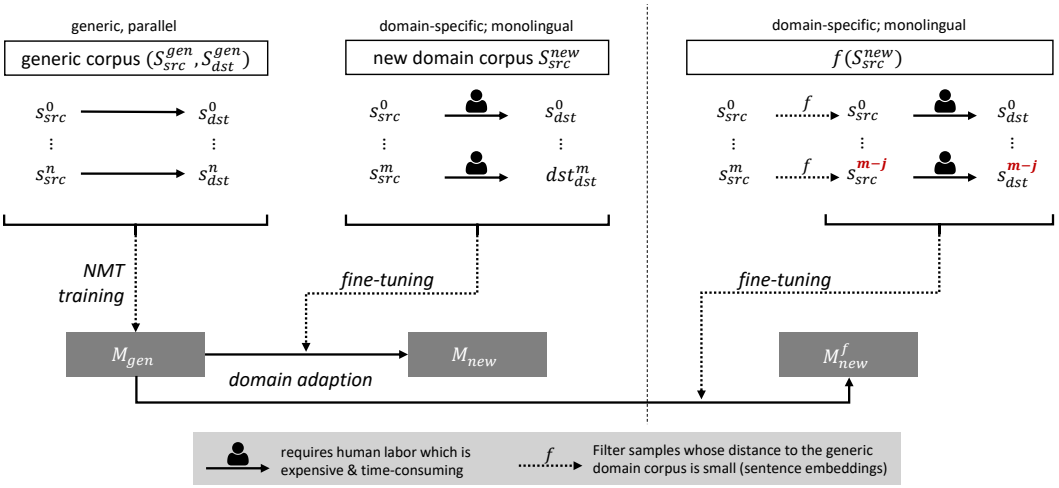


Figure 1: We focus on domain adaptation of a generic NMT model M_{gen} with humans-in-the-loop that translate monolingual data of the new domain with limited resources. We simulate crowd-translated content using two parallel corpora S^{new} representing data of new domains for training the adapted model M_{new}. We propose an advanced query strategy for selecting sentences from S^{new} that need to be translated by their similarity to the generic corpus.

line, rare n-gram sampling, and a sampling based on word confidences. In a recent work, Lam et al. (2018) suggest to incorporate human judgments on partial translations as reinforcement signal for improving NMT models and evaluate it in a simulation experiment with existing parallel corpora. For reducing human workload, they suggest an entropy-based method to trigger human judgments similar to active learning approaches with human oracles.

We focus on a query strategy for domain adaptation of NMT models based on active learning. We consider settings in which human workers provide new training data (Barz et al., 2018b,a; Green et al., 2015) for domains with no or only little parallel corpora due to, for instance, budget constraints. Our experiment includes random sampling as a baseline similar to González-Rubio et al. (2012) and an advanced sentence selection strategy based on distances between sentence embeddings that also encode the semantics of a sentence (Arora et al., 2017), adapted for Arabic.

3 Method

We implement a baseline query strategy (random sampling) and an advanced query strategy (see 3.3) for selecting training samples which are used for fine-tuning a generic NMT model. In this section, we describe the applied NMT model and the generic training process, as well as the two query strategies used in the domain adaptation process.

3.1 Model Architecture and Training

We use the TensorFlow implementation of NMT[2] (Luong et al., 2017) configured as an 8-layered bidirectional RNN with standard LSTM cells in each layer and residual connections between the layers. We use the same architecture for both, generic model training and fine-tuning tasks. The model is trained[3] with vanilla SGD for $350k$ iterations with a batch size of 50 and a dropout rate of 0.2. The initial learning rate is set to 1.0 and a decay factor of 0.5 is applied after every $1k$ iterations starting from $170k$ iterations. We use the standard hyperparameters provided in the NMT framework and set the vocabulary size to $32k$ for both Arabic and English. We train the generic model (M_{gen}) for one week using the LDC corpus (S^{gen}) (see Figure 1) and use the resulting checkpoint for all of our fine-tuning experiments.

3.2 Datasets and Preprocessing

In our experiment, we use the LDC Newswire corpus (Munteanu and Marcu, 2005) and two publicly available datasets, MEDAR and GlobalVoices. The corpus statistics are summarized in Table 1. The LDC Newswire parallel corpus (Ar-En) is used for training a *generic* model and the MEDAR and GlobalVoices datasets for *domain-specific* fine-tuning. We include datasets on two

[2] https://github.com/tensorflow/nmt

[3] All experiments were performed on an Ubuntu machine (Intel i7-5960X) with 8 cores and 2 GTX-1080 graphics cards

Corpus	Sentence Pairs	Domain	Usage
LDC Newswire	1.3M	Generic	Generic model training
MEDAR	0.5k	Climate Change	Domain specific fine-tuning
GlobalVoices	37k	Politics, Human Rights	Domain specific fine-tuning

Table 1: Details of datasets that we used in our experiments.

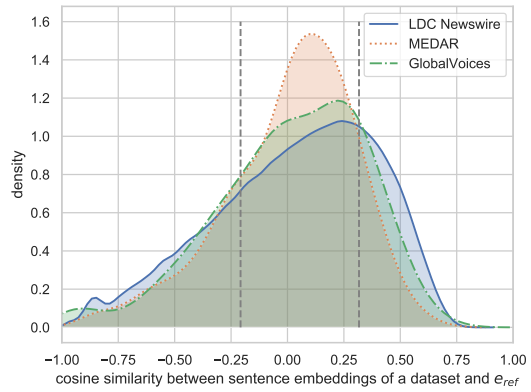


Figure 2: Kernel density estimates for the distributions of distances (cosine similarity) between sentence embeddings of each considered dataset and e_{ref}, the mean of sentence embeddings of the dataset used for training of the generic translation model. The gray dotted lines represent the 25% and the 75% percentile of the distance distribution for the generic model.

different genres to investigate whether our findings generalize irrespective of the domain of the fine-tuning set. To reduce noise in the data, we clean the datasets by discarding instances with mixed tokens (i.e. English sentences containing Arabic words or Arabic sentences containing English words). This step removes around 0.01%, 1.2%, and 10.14% of sentence pairs from LDC Newswire, MEDAR, and GlobalVoices datasets respectively. Further preprocessing steps of our system pipeline include normalization and tokenization[4] of the sentences and generation of *byte pair encodings* (BPE)[5] (Sennrich et al., 2016) for the tokenized sentences and the vocabulary.

3.3 Query Strategies for Sample Selection

For model adaptation in limited resource settings, it is desirable to reduce the number of samples from the target domain and, thus, the required time and cost for receiving human translations. Our goal is to develop a query strategy for selecting the most informative update samples, similar to the active learning paradigm. We propose a method that estimates the informativeness of a sample based on its similarity to the generic corpus using sentence embeddings. We exclude semantically overlapping parts from the new corpus which reduces the amount of training samples that need to be translated by human labor and that need to be included in model training (see Figure 1). We refer to this method as fine-tuning with *advanced sampling*. In addition, we implement a baseline method which selects all samples from a new domain in random order (fine-tuning with *random sampling*).

For our advanced sampling method, we use *smooth inverse frequency* (SIF)-based sentence embeddings (Arora et al., 2017) extended for Arabic which we refer to as AraSIF (see Section 3.4). It encodes sentences from the source language $s \in S_{src}$ into a 300-dimensional vector e_s:

$$e_{sif} : S_{src} \rightarrow \mathbb{R}^{300}, s \mapsto e_s$$

Arora et al. (2017) show that SIF-based embeddings perform well for many semantic textual similarity tasks. This implies that the sentences which are close to each other in the embedding space can be considered to be semantically similar. We estimate the informativeness of a sample for domain adaptation based on the semantic similarity of two sentences. We use the cosine distance d between two sentence embeddings e_s and $e_{s'}$ as a proxy for semantic similarity:

$$d : \mathbb{R}^{300\times 2} \rightarrow [-1, 1], (e_s, e_{s'}) \mapsto d_{s,s'}$$

Hereby, the mean of all sentence embeddings of the generic corpus $S^{gen}_{src} \subset S_{src}$ serves as the reference point e_{ref} in the sentence embedding space for comparing sentences from other corpora:

$$e_{ref} = mean(e_{sif}(S^{gen}_{src})), e_{ref} \in \mathbb{R}^{300}$$

Calculating the cosine similarity between all samples of a new domain $S^{new}_{src} \subset S_{src}$ and this

[4]https://github.com/moses-smt/mosesdecoder
[5]https://github.com/rsennrich/subword-nmt

reference point, results in a distribution of distances indicating the semantic similarity or dissimilarity of samples from the new domain to the generic domain. We show the distributions for all considered datasets in terms of a kernel density estimate in Figure 2: MEDAR and GlobalVoices as new domains and LDC Newswire as generic reference domain. Initially, we anticipated the target domain corpora to partially lie outside of the reference distribution, but the new corpora rather seem to be more specific subsets of the generic domain. Therefore, we select training samples for our fine-tuning process from the new domains that belong to the long-tail of the distance distribution of the generic domain corpus. We expect the informativeness of a new sample s to be high, if it is underrepresented in the generic domain dataset in terms of semantic similarity to e_{ref}, this is if $d_{s,e_{ref}}$ is high. The interval boundaries that frame the longtail are the only parameters that need to be defined for this approach. We use the 25% and 75% percentiles of the distance distribution of the generic domain to define these outer regions (see dotted vertical lines in Figure 2).

3.4 AraSIF: Arabic Sentence Embeddings

To obtain sentence embeddings for Arabic sentences we propose AraSIF. We use SIF[6] with AraVec[7] (Soliman et al., 2017), a Word2Vec pretrained model that is trained on 1.8M Arabic Wikipedia articles with a total vocabulary size of 662k. SIF is based on word weights for computing embeddings, for which we consider all tokens with a frequency count of at least 200. We preprocess the Wikipedia articles on which AraVec was trained on, for computing the word frequency. In addition, SIF expects the word embedding to be in GloVe embedding format. Hence, we convert the AraVec word embeddings from Word2Vec to GloVe format. The code for AraSIF is publicly available at *DFKI Interactive Machine Learning* repository on GitHub.

4 Experiment

We conduct a simulation experiment for investigating the effectiveness of our advanced query strategy in reducing the required amount of update samples for adapting an NMT model to a new domain. Our approach selects samples using monolingual information only, which can be assumed to be available without investing resources. For this, we compare the translation quality when fine-tuning a model with *random sampling* and when fine-tuning it with a reduced number of update samples resulting from our *advanced sampling*. Our generic NMT model M_{gen} is adapted to two new domains, represented by the GlobalVoices and the MEDAR datasets, using both query strategies. We include a varying number of epochs for identifying good training parameters. We hypothesize that our advanced query strategy for sample selection can effectively reduce the number of fine-tuning samples without hampering the translation quality when compared to the baseline.

epochs	1	5	10	20
n_t	T_{sgd}	T_{sgd}	T_{sgd}	T_{sgd}
50	1	5	10	20
100	2	10	20	40
150	3	15	30	60
200	4	20	40	80
250	5	25	50	100
300	6	30	60	120
350	7	35	70	140
400	8	40	80	160

Table 2: Considered combinations of update set sizes (n_t) and SGD updates or iterations (T_{sgd}) used for fine-tuning.

5 Evaluation Procedure

We perform the simulation experiment with two new domain datasets and observe the impact of different parameters on domain adaptation of the generic NMT model using small amounts of new samples. These can be considered to stem from human workers, e.g., professional translators or crowdworkers. Considered parameters include the number of training samples in the update set n_t and the number of training epochs e. The number of epochs defines the number of training iterations: The number of Stochastic Gradient Descent (SGD) updates, denoted by T_{sgd}, is computed by:

$$T_{sgd} = \left\lceil \frac{n_t}{|S|} \right\rceil \cdot e$$

where $|S|$ is the mini-batch size which we set to 50 throughout our experiments, n_t is the number of sentence pairs in the *update set*, and *e* is the number of epochs. Table 2 provides an overview of all considered configurations.

[6] https://github.com/PrincetonML/SIF
[7] https://github.com/bakrianoo/aravec

The update sets used for model adaptation are generated from either MEDAR or GlobalVoices dataset, after excluding a static test set of 100 sentences for each. Both update sets are constrained to a maximum sample size of 400 to allow a fair comparison (this is the maximum size for MEDAR, see table 1). Further, we assume that the amount of data from the new domain might be small due to resource constraints or scarcity.

For the *random sampling* case, we select all 400 samples from each of the datasets in a random order and use it to adapt our generic model for all parameter configurations in Table 2. Samples 1 to 50 of the update set are used for training the $n_t = 50$ model for all epochs. The model fine-tuning is continued with samples 51 to 100 for the $n_t = 100$ model for all epochs, and so on. Considering the stochastic nature of SGD, we repeat the experiment 5 times and report the average scores on the respective test sets, instead of providing a point estimate. We observe the training times on the update set and the BLEU scores (Papineni et al., 2002) on the test set as a dependent variable.

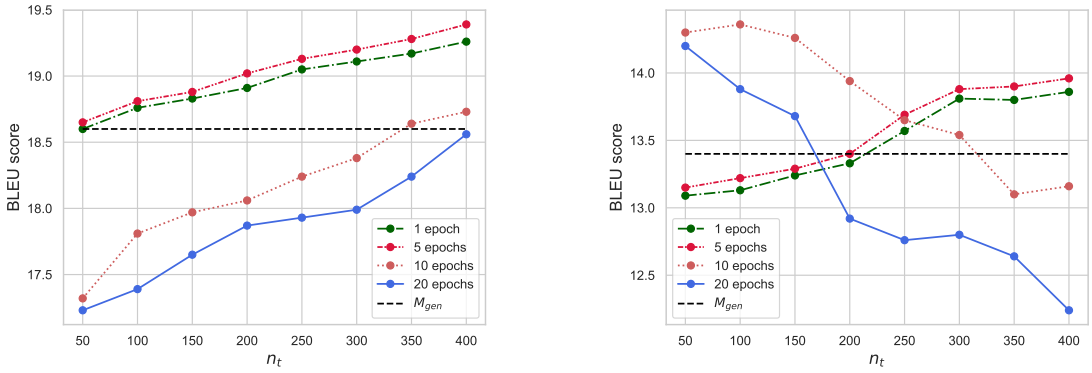
For our *advanced sampling* strategy, we select a subset of all training samples for both datasets using the filter mechanism described above. We include a sentence s, if its cosine distance d to e_{ref} is smaller than the 25% percentile (-0.208) or larger than the 75% percentile (0.317) of the generic distance distribution (see Figure 2). This leaves us with 135 fine-tuning samples for MEDAR and 169 for GlobalVoices from the original 400 samples. We consider the same set of parameters than before with the difference that the size of the update set n_t is limited to the reduced number of samples.

5.1 Results

In this section, we present the results of our fine-tuning experiments for adapting the generic model with different sampling strategies. We use an increasing number of update samples (n_t), different epoch configurations (e) and two new domain datasets. The generic baseline model M_{gen} achieves a reference BLEU score of 18.6 for MEDAR and 13.4 for GlobalVoices. We used the same test set which we have used for evaluating the fine-tuned model.

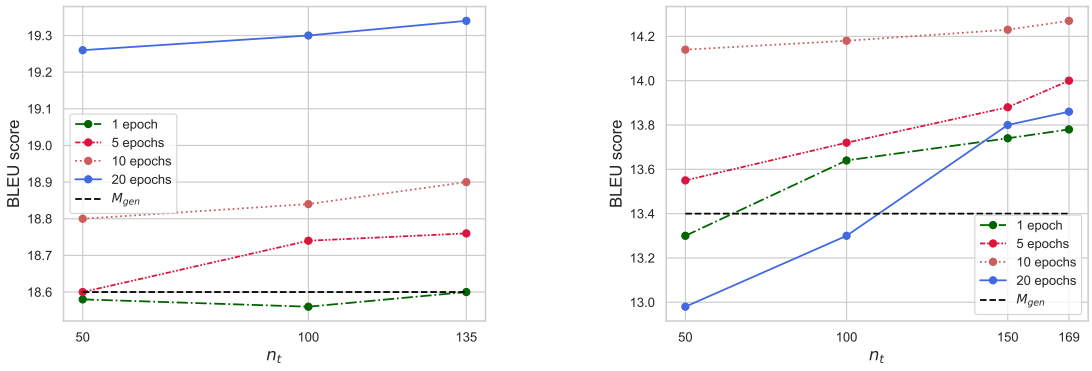
Figure 3 summarizes the BLEU scores for all parameter settings and both new domains concerning the *random sampling* condition. For MEDAR, we can observe a monotonic improvement in BLEU score for increasing numbers of samples n_t in the update set for all epoch configurations. However, compared to the reference score of 18.6, only $e = 1$ and $e = 5$ achieve meaningful improvements: we can observe an improvement after fine-tuning with first two mini-batches. Higher numbers for e (10, 20) result in lower BLEU scores than the reference, also when including all samples ($n_t = 400$). Only for $e = 10$ and $n_t = [350, 400]$ we observe a BLEU score slightly better than the reference model. The best BLEU score on the MEDAR test set is achieved using $n_t = 400$ and $e = 5$ with a score of 19.39. Averaged over 5 repetitions of the experiment, the runtime ranges between $59s$ for $n_t = 50$ and $68s$ for $n_t = 400$ for $e = 1$. All other configurations require longer training times. For GlobalVoices, we observe a monotonic improvement with increasing number of samples n_t for $e \in \{1, 5\}$. Higher numbers for the epoch configuration result in a monotonic deterioration of BLEU score. In contrast to the models adapted to MEDAR, training with $e \in \{10, 20\}$ yields better results than the reference score of 13.4 for small n_t. Yet, due to the negative trend in BLEU scores, models with these epoch configurations fall below the reference score. The best BLEU score on the GlobalVoices test set is achieved using $n_t = 100$ and $e = 10$ with 14.36. Averaged over 5 trainings, the runtime ranges between $62s$ for $n_t = 50$ and $122s$ for $n_t = 400$. For $n_t \geq 200$, we observe better BLEU scores than the reference model for $e = 1$ and $e = 5$, where the training times for $e = 1$ grow considerably slower than for $e = 5$. Here, the training times range between $49s$ for $n_t = 50$ and $66s$ for $n_t = 400$.

Figure 4 summarizes the BLEU scores for all considered settings and domains for our *advanced sampling* condition. For MEDAR, we observe improvements in BLEU scores similar to the random sampling condition. All epoch configurations, except for $e = 1$, achieve scores higher than the reference (18.6) starting from the first update set. With the advanced sample selection, the best score of 19.34 is achieved using $n_t = 135$ and $e = 20$. Using $e = 1$ for varying number of samples (n_t) yields BLEU scores which are slightly lower than the reference BLEU score of 18.6. Averaged over 5 trainings, the runtime ranges between $54s$ and $58s$ for $e = 1$ and between $59s$



(a) MEDAR (b) GlobalVoices

Figure 3: BLEU scores of fine-tuned NMT models for MEDAR and GlobalVoices corpora with *random sampling* for varying sizes of the update set (n_t) and different number of training epochs (e).



(a) MEDAR (b) GlobalVoices

Figure 4: BLEU scores of fine-tuned NMT models for MEDAR and GlobalVoices corpora with *advanced sampling* for varying sizes of the update set (n_t) and different number of training epochs (e).

and $70s$ for $e = 5$. For GlobalVoices, we obtain the best score of 14.27 with $e = 10$ which is comparable to $e = 10$ and $n_t = 100$ in the *random sampling* condition. For $e \in \{5, 10\}$, we observe better scores compared to the random sampling condition, for all update set sizes n_t. In addition, with our advanced sampling, we always observe a monotonic increase in BLEU score for all epoch configurations and increasing number of samples in the update set n_t, in contrast to the epoch configurations $e \in \{10, 20\}$ for the random sampling condition where we observe a decreasing trend in BLEU scores. The runtimes are similar to training times of MEDAR models.

6 Discussion

Our experiment shows that fine-tuning the generic model M_{gen} with *random sampling* for small update sets can improve BLEU scores (see Figure 3). In particular, we observe improvements over the baseline with MEDAR data for $e = \{1, 5\}$ and with GlobalVoices data for $e = \{1, 5\}$ for update set sizes larger than 200 and for $e = \{10, 20\}$ with update set sizes less than 200. We did not find *log-like* relations similar to Barone et al. (2017). The reason for this could be because we included less data for the domain adaptation. For the random sampling condition, with MEDAR dataset, we can trade translation quality for faster training times since $e = 5$ training yields only slightly better BLEU scores when compared to $e = 1$ setting. Analogously, for GlobalVoices dataset, $e = 1, 5$ achieves similar performance and perform better than baseline model when $n_t > 200$, which allows to switch to a faster model training with $e = 1$ with a marginal loss in translation quality. Con-

cerning larger values of e for both the new domains yield a slower gain in translation quality or even a loss in translation quality for $e \in \{10, 20\}$ (GlobalVoices) after an initial improvement over the baseline. This loss might be caused by overfitting to the training samples due to a high number of training iterations.

Using our *advanced* sampling for fine-tuning M_{gen} to a new domain, significantly reduces the amount of training samples without loss in translation quality compared to the commonly used fine-tuning with *random sampling*. This allows to dramatically reduce the amount of data that needs to be translated or post-edited by human labor, because the sampling of "unlabeled" instances is performed using monolingual data only. In case of MEDAR, our method reduces translation cost and time by 66.25% compared to random sampling. In addition, the BLEU scores improved overall: Except for $e = 1$ training setting, none of the scores is lower than the baseline score. An interesting observation when compared to the random sampling condition is that samples resulting from our advanced sampling need more epochs to achieve better BLEU scores. We believe this is due to the following two reasons: (i) Domain mismatch: the genre of samples of MEDAR dataset is significantly different from the domain of the samples observed in M_{gen} (Almahairi et al., 2016). (ii) Low amount of samples: our advanced sampling approach removes 66% of samples from the original 400. Both of these factors necessitates more training epochs to achieve the best BLEU score as with random sampling condition. In case of GlobalVoices, we can observe similar improvements in BLEU score for $e = \{5, 10\}$: we achieve a similar BLEU score as with random sampling baseline although we excluded 57.75% of the training data. All in all, we can confirm our hypothesis that our advanced sampling query strategy for sample selection effectively reduces the number of fine-tuning samples without degrading the translation quality compared to results of the baseline. A further advantage of our approach is that it supports continuous fine-tuning, in contrast to other methods which require a complete re-training of the model whenever new samples of the target domain become available (Wang et al., 2017).

Currently, there is one limitation in our work: The update sets in our evaluation are quite small. Hence, we want to investigate the performance of our method using all 36k samples of the GlobalVoices parallel corpus.

A promising direction for future work would be to investigate the impact of active learning in NMT using our advanced sentence sampling on translation time and quality of incremental model improvements. In settings with human workers that post-edit translation candidates, translations that improve over time might reduce this post-editing effort and, consequently reduce the overall time and budget required for model adaptation to a new domain. In addition, this technology can increase the efficiency of ubiquitous machine translation interfaces, e.g., for multimodal post-editing (Herbig et al., 2019; Oviatt et al., 2017), real-time translation systems in virtual reality (Toyama et al., 2014), or medical cross-language dialogue applications (Sonntag et al., 2009b,a) As a follow-up work, we would like to experiment with a *clustering-based sample selection* instead of using a single reference vector (e_{ref}) for the whole generic domain and observe the performance of domain-adapted sequence-to-sequence models based on the chosen samples.

7 Conclusion

We investigate the problem of incremental domain adaptation of a generic NMT model in a limited resources setting. Our NMT models improve BLEU score with only small amounts of data from a new domain. Hereby, sentences from the source language were randomly sampled for being used as parallel training data after human translations. We simulated the human translation task by using existing parallel corpora. Also, we introduced an *advanced sampling* strategy, based on semantic text similarity using a state-of-the-art technique, after extending it for computing sentence embeddings for Arabic (AraSIF). We found that our novel method achieves similar BLEU scores, compared to fine-tuning with random sampling, but using less than half of the initial training data. This enables more efficient domain adaptation of NMT models with humans-in-the-loop and with resource constraints.

Acknowledgments

This work is supported by EIT Digital (H2020).

References

Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron C. Courville. 2016. First result on arabic neural machine translation. *CoRR*, abs/1606.02680.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings.

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494, Copenhagen, Denmark. Association for Computational Linguistics.

Michael Barz, Neslihan Büyükdemircioglu, Rikhu Prasad Surya, Tim Polzehl, and Daniel Sonntag. 2018a. Device-Type Influence in Crowd-based Natural Language Translation Tasks (short paper). In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD) 2018 and CrowdBias 2018) co-locate*, volume 2276 of *CEUR Workshop Proceedings*, pages 93–97. CEUR-WS.org.

Michael Barz, Tim Polzehl, and Daniel Sonntag. 2018b. Towards hybrid human-machine translation services. EasyChair Preprint no. 333.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 385–391. Association for Computational Linguistics.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254. Association for Computational Linguistics.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2015. Natural Language Translation at the Intersection of AI and HCI. *Queue*, 13(6):30.

Nico Herbig, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019. Multi-modal approaches for post-editing machine translation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 231:1–231:11, New York, NY, USA. ACM.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tsz Kin Lam, Julia Kreutzer, and Stefan Riezler. 2018. A reinforcement learning approach to interactive-predictive neural machine translation. *CoRR*, abs/1805.01553.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126. Association for Computational Linguistics.

Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural machine translation (seq2seq) tutorial. *https://github.com/tensorflow/nmt*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31(4):477–504.

Laurent Nepveu, Guy Lapalme, Philippe Langlais, and George Foster. 2004. Adaptive language and translation models for interactive machine translation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Sharon Oviatt, Björn Schuller, Philip R Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger. 2017. Introduction: Scope, Trends, and Paradigm Shift in the Field of Computer Interfaces. In *The Handbook of Multimodal-Multisensor Interfaces*, pages 1–15. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA.

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 - ACL '02*, page 311, Morristown, NJ, USA. Association for Computational Linguistics.

Álvaro Peris and Francisco Casacuberta. 2018. Online learning for effort reduction in interactive neural machine translation. *CoRR*, abs/1802.03594.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Burr Settles. 2010. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R. El-Beltagy. 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256 – 265. Arabic Computational Linguistics.

Daniel Sonntag, Robert Nesselrath, Gerhard Sonnenberg, and Gerd Herzog. 2009a. Supporting a rapid dialogue system engineering process. *Proceedings of the 1st IWSDS*.

Daniel Sonntag, Pinar Wennerberg, Paul Buitelaar, and Sonja Zillner. 2009b. Pillars of ontology treatment in the medical domain. *J. Cases on Inf. Techn.*, 11(4):47–73.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3104–3112. MIT Press.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).

Takumi Toyama, Daniel Sonntag, Andreas Dengel, Takahiro Matsuda, Masakazu Iwamura, and Koichi Kise. 2014. A mixed reality head-mounted text translation system using eye gaze input. In *Proceedings of the 19th international conference on Intelligent User Interfaces - IUI '14*, pages 329–334, New York, New York, USA. ACM Press.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 560–566. Association for Computational Linguistics.