Non-Parametric Unsupervised Domain Adaptation for Neural Machine Translation

Xin Zheng[1], Zhirui Zhang[2], Shujian Huang[13]*, Boxing Chen[2], Jun Xie[2], Weihua Luo[2] and Jiajun Chen[1]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Language Technology Lab, Alibaba DAMO Academy [3]Peng Cheng Laboratory, China
[1]zhengxin@smail.nju.edu.cn, {huangsj,chenjj}@nju.edu.cn
[2]{boxing.cbx, qingxing.xj, weihua.luowh}@alibaba-inc.com
[2]zrustc11@gmail.com

Abstract

Recently, kNN-MT (Khandelwal et al., 2020) has shown the promising capability of directly incorporating the pre-trained neural machine translation (NMT) model with domain-specific token-level k-nearest-neighbor (kNN) retrieval to achieve domain adaptation without retraining. Despite being conceptually attractive, it heavily relies on high-quality in-domain parallel corpora, limiting its capability on unsupervised domain adaptation, where in-domain parallel corpora are scarce or nonexistent. In this paper, we propose a novel framework that directly uses in-domain monolingual sentences in the target language to construct an effective datastore for k-nearest-neighbor retrieval. To this end, we first introduce an autoencoder task based on the target language, and then insert lightweight adapters into the original NMT model to map the token-level representation of this task to the ideal representation of translation task. Experiments on multi-domain datasets demonstrate that our proposed approach significantly improves the translation accuracy with target-side monolingual data, while achieving comparable performance with back-translation. Our implementation is open-sourced at https://github.com/zhengxxn/UDA-KNN.

1 Introduction

Non-parametric methods (Gu et al., 2018; Zhang et al., 2018a; Bapna and Firat, 2019a; Khandelwal et al., 2020; Zheng et al., 2021) have recently been successfully applied to neural machine translation (NMT). These approaches complement advanced NMT models (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017; Hassan et al., 2018) with external memory to alleviate the performance degradation when translating out-of-domain sentences, rare words (Koehn and Knowles, 2017), etc. Among them, kNN-MT (Khandelwal et al., 2020) is a simple yet effective non-parametric method using nearest neighbor retrieval. More specifically, kNN-MT equips a pre-trained NMT model with a kNN classifier over a provided datastore of cached context representations and corresponding target tokens to improve translation accuracy without retraining. This promising ability to access any provided datastore or external knowledge during inference makes it expressive, adaptable, and interpretable.

Despite the potential benefits, kNN-MT requires large-scale in-domain parallel corpora to achieve domain adaptation. However, in practice, it is not realistic to collect large amounts of high-quality parallel data in every domain we are interested in. Since monolingual in-domain data is usually abundant and easy to obtain, it is essential to explore the capability of kNN-MT on unsupervised domain adaptation scenario that utilizes large amounts of monolingual in-domain data. One straightforward and effective solution for unsupervised domain adaptation is to build in-domain synthetic parallel data via back-translation of monolingual target sentences (Sennrich et al., 2016a; Zhang et al., 2018b; Dou et al., 2019; Wei et al., 2020). Although this approach has proven superior effectiveness in exploiting monolingual data, applying it in kNN-MT requires an additional reverse model and brings the extra cost of generating back-translation, making the adaptation of kNN-MT more complicated and time-consuming in practice.

In this paper, we propose a novel **U**nsupervised **D**omain **A**daptation framework based on kNN-MT (UDA-kNN). The UDA-kNN aims at directly leveraging the monolingual target-side data to generate the corresponding datastore, and encouraging it to play a similar role with the real bilingual in-domain data, through the carefully designed architecture and loss function. Specifically, we introduce an autoencoder task based on target language to enable datastore construction with monolingual data.

* Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4234–4241
November 7–11, 2021.

Then we incorporate lightweight adapters into the encoder part of pre-trained NMT model to make the decoder's representation in autoencoder task close to the corresponding representation in translation task. In this way, the adapter module implicitly learns the semantic mapping from the target language to source language in feature space to construct an effective in-domain datastore, while saving the extra cost of generating synthetic data via back-translation.

We evaluate the proposed approach on multi-domain datasets, including IT, Medical, Koran and Law domains. Experimental results show that when using target-side monolingual data, our proposed approach obtains 4.9 BLEU improvements on average and even achieves similar performance compared with back-translation.

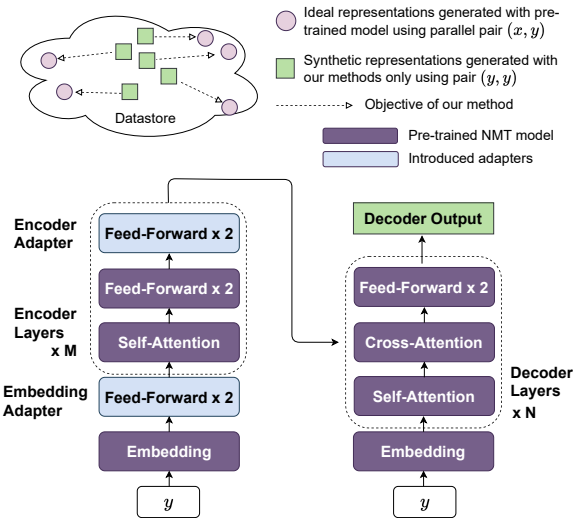

Figure 1: An overview of the proposed method.

2 Background

In this section, we give a brief introduction to the domain adaptation of kNN-MT. In general, the process includes two steps: creating an in-domain datastore and decoding with retrieval on it.

In-domain Datastore Creation. Given a pre-trained general domain NMT model and an in-domain parallel corpus, kNN-MT utilizes the model to forward pass the corpus to create a datastore. Formally, for each bilingual sentence pair in the corpus $(x, y) \in (\mathcal{X}, \mathcal{Y})$, the NMT model will generate a context representation $h(x, y_{<t})$ for each target-side token y_t. Then, the datastore is constructed by collecting the representations and corresponding tokens as keys and values respectively:

$$(\mathcal{K}, \mathcal{V}) = \bigcup_{(x,y)\in(\mathcal{X},\mathcal{Y})} \{(h(x, y_{<t}), y_t), \forall y_t \in y\}. \tag{1}$$

Decoding with Retrieval. On each decoding step t, the NMT model first generates a representation $h(x, \hat{y}_{<t})$ for current translation context, which consists of source-side x and generated target-side tokens $\hat{y}_{<t}$. Then, the representation is used to query the in-domain datastore for k nearest neighbors, which can be denoted as $N = \{(h_i, v_i), i \in \{1, 2, ..., k\}\}$. These neighbors are utilized to form a distribution over the vocab:

$$p_{\text{kNN}}(y_t|x, \hat{y}_{<t}) \propto \sum_{(h_i,v_i)} \mathbb{1}_{y_t=v_i} \exp\left(\frac{-d(h_i, h(x, \hat{y}_{<t}))}{T}\right), \tag{2}$$

where T is the temperature and $d(\cdot, \cdot)$ indicates the squared euclidean distance. The final probability to predict next token $y_{<t}$ is an interpolation of two distributions with a hyper-parameter λ:

$$\begin{aligned} p(y_t|x, \hat{y}_{<t}) = \lambda\, & p_{\text{kNN}}(y_t|x, \hat{y}_{<t}) \\ & + (1-\lambda)\, p_{\text{NMT}}(y_t|x, \hat{y}_{<t}), \end{aligned} \tag{3}$$

where p_{NMT} indicates the general domain NMT prediction and p_{kNN} represents the in-domain retrieval based prediction.

3 Unsupervised Domain Adaptation with kNN-MT

Although Khandelwal et al. (2020) has shown the capability of kNN-MT on domain adaptation, the datastore creation heavily relies on high-quality in-domain parallel data, which cannot be always satisfied in practice. As in-domain monolingual data is usually abundant and easy to obtain, it is essential to extend the capability of kNN-MT on unsupervised domain adaptation that merely uses large amounts of in-domain target sentences. In this paper, we design a novel non-parametric **U**nsupervised **D**omain **A**daptation framework based on kNN-MT (UDA-kNN) to fully leverage in-domain target-side monolingual data.

The overview framework of UDA-kNN is illustrated in Figure 1. The UDA-kNN starts with the autoencoder task based on target language y, where the target-side is simply copied to the source-side to generate pair (y, y). Based on that, the UDA-kNN aims to make the decoder's representation in autoencoder task close to the ideal representation in

translation task. In this way, we can directly leverage autoencoder structure and in-domain target sentences to construct the corresponding datastore for k-nearest-neighbor retrieval, which is similar to that from real in-domain bilingual data. Next, we will introduce the architecture and training objective of our proposed method in detail.

Architecture. We insert lightweight adapter layers (Houlsby et al., 2019; Guo et al., 2020, 2021) into the source embedding layer and each encoder layer of the pre-trained NMT model to perform the autoencoder task, by which we only increase a few parameters for our method. Specifically, we simply construct the adapter layer with layer normalization as well as two feed-forward networks with non-linearity between them:

$$Z = W_1 \cdot (\mathrm{LN}(H)), H_{\mathrm{o}} = H + W_2 \cdot (\mathrm{ReLU}(Z)), \tag{4}$$

where H and H_{o} are the input and output hidden states of the adapter layer respectively, LN indicates layer normalization, W_1 and W_2 are the parameters of the feed-forward networks.

Training. The UDA-kNN is designed to leverage monolingual target-side data to generate the corresponding datastore, which plays a similar role with real in-domain bilingual data. We achieve this by leveraging out-of-domain bilingual data $(\mathcal{X}, \mathcal{Y})$. More specifically, given a bilingual sentence pair in the corpus $(x, y) \in (\mathcal{X}, \mathcal{Y})$, the original NMT model generates decoder representation $h_{(x;\, y_{<t})}$ for each target token y_t. Meanwhile, with the target-copied pair (y, y), the NMT model with adapters generates another representation for each y_t, which can be denoted as $h'_{(y;\, y_{<t})}$. We take the end-to-end paradigm to directly optimize the adapter layers by minimizing the squared euclidean distance of the two sets of decoder representations:

$$\theta^* = \min_{\theta} \sum_{(x,\, y) \in (\mathcal{X}, \mathcal{Y})} \sum_{t} ||h'_{(y;\, y_{<t})} - h_{(x;\, y_{<t})}||^2, \tag{5}$$

where θ is the parameters of all adapter layers. Note that we keep original parameters in the pre-trained NMT model fixed during training to avoid the performance degradation of the NMT model in the inference stage.

Prediction. For unsupervised domain adaptation, given the domain-specific target-side monolingual data, we first copy the target sentences to the source side to generate synthetic bilingual pairs. Then the pre-trained NMT model with adapter layers forward passes these pairs to create an in-domain datastore. When translating in-domain sentences, we utilize the original NMT model and kNN retrieval on the in-domain datastore to perform online domain adaptation as Equation (3).

4 Experiments

4.1 Setup

Datasets and Evaluation Metric. We use the same multi-domain dataset as Aharoni and Goldberg (2020) to evaluate the effectiveness of our proposed model and consider domains including **IT**, **Medical**, **Koran**, and **Law** in our experiments. We extract target-side data in the training sets to perform unsupervised domain adaptation while keeping the dev and test sets unchanged. Besides, WMT'19 News data[1] is used for training the adapters in our method as well as the reverse translation model for back-translation. The sentence statistics of all datasets are illustrated in table 1. The Moses toolkit[2] is used to tokenize the sentences and we split the words into subword units (Sennrich et al., 2016b) with the codes provided by the pre-trained model (Ng et al., 2019). We use SacreBLEU[3] to measure all results with case-sensitive detokenized BLEU (Papineni et al., 2002).

Dataset	WMT19'News	IT	Medical	Koran	Laws
Train	37,079,168	222,927	248,009	17,982	467,309
Dev	10,000	2000	2000	2000	2000
Test	-	2000	2000	2000	2000

Table 1: Statistics of dataset in different domains.

Methods. We compare our proposed approach with several baselines:

- **Basic NMT**: A general domain model is directly used to evaluate on in-domain test sets.
- **Empty-kNN**: The source-side of synthetic bilingual data is always set to *<EOS>* token.
- **Copy-kNN**: Each target sentence is copied to source-side to produce synthetic bilingual data. This is a special case of our method without model training.
- **BT-kNN**: A reverse translation model is applied to produce synthetic bilingual data, which are used to generate in-domain datastore.

[1] http://www.statmt.org/wmt19/translation-task.html

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

[3] https://github.com/mjpost/sacrebleu

Model	IT	Medical	Law	Koran	Avg.
Basic NMT	38.35	39.99	45.48	16.26	35.02
Empty-kNN	38.06	40.01	45.62	16.44	35.03
Copy-kNN	38.96	40.86	46.00	17.06	35.72
BT-kNN	41.35	47.02	52.91	19.58	40.23
UDA-kNN	41.57	46.64	52.02	19.42	39.91
Parallel-kNN	45.96	54.16	61.31	20.30	45.43

Table 2: The BLEU scores [%] of different methods evaluated on four domains.

- **Parallel-kNN**: Ground-truth parallel data is used to generate the in-domain datastore, which can be regarded as the upper bound of the kNN retrieval based methods.

Implementation Details. We use the WMT'19 German-English News translation task winner model (Ng et al., 2019) as our general domain model. For introduced adapters, the hidden size is set to 1024, with only about 6% parameters of the original model. Adam (Kingma and Ba, 2015) is used to update the parameters in adapters. During training, we collect about 40000 tokens for each batch and schedule the learning rate with the inverse square root decay scheme, in which the warm-up step is set as 4000, and the maximum learning rate is set as 7e-4. Faiss[4] is used to build the in-domain datastore to carry out fast nearest neighbor search. We utilize faiss to learn 4096 cluster centroids for each domain, and search 32 clusters for each target token in decoding. When inference, we retrieve 16 nearest neighbors in the datastore. We set the hyper-parameter T as 4 for IT, Medical, Law, and 40 for Koran. The λ is tuned on the in-domain dev sets for different methods.

4.2 Main Results

The adaptation performance of different methods are listed in Table 2. Obviously, our method can significantly improve the translation accuracy on in-domain test sets compared to basic NMT, while Empty-kNN and Copy-kNN can't. It demonstrates the efficiency of our proposed method to create an in-domain datastore by leveraging only monolingual data. Besides, we can observe that our method achieves comparable performance over BT-kNN, but completely avoids the reverse model and extra time cost to generate synthetic data, making the adaptation much faster and simpler.

[4] https://github.com/facebookresearch/faiss

4.3 Analysis

In this section, we would like to further explore the reasons behind the success of our approach.

Similarity Measurement. We measure the cosine similarity and squared euclidean distances between the synthetic representations generated by our method and ideals generated using ground-truth parallel data. As shown in Table 3, we also list the results of BT-kNN and Copy-kNN. We can observe that even without the source language information, our UDA-kNN can generate the representations that are close enough to the ideals as BT-kNN, leading to the efficient in-domain retrieval for kNN-MT. It also verifies the effectiveness of the adapter layers on directly learning the semantic mapping from target language to source language in feature space.

Cosine Similarity (↑)				
Method	IT	Medical	Law	Koran
Copy-kNN	0.74	0.77	0.77	0.65
BT-kNN	0.85	0.86	0.92	0.81
UDA-kNN	0.87	0.87	0.91	0.84
Squared Euclidean Distance (↓)				
Method	IT	Medical	Law	Koran
Copy-kNN	85.93	79.03	77.97	145.93
BT-kNN	47.00	42.30	25.47	78.77
UDA-kNN	46.10	45.83	31.36	68.56

Table 3: Cosine similarity / squared euclidean distance between the ground-truth representations and that generated by different methods.

Visualization. We also collect and visualize the representations with the same target tokens in different datastores to give intuitive insights of the impact of adapters. Specifically, we select three common words in IT domain and show the results in Figure 2. We can see that the representations generated with Copy-kNN tend to gather in small areas, which results in retrieval collapse when meeting diverse translation contexts. While with the adapters, the distribution of the same label in the datastore can be closer to that generated with bilingual pairs, improving the retrieval efficiency.

Effect of Adapter Position. In our proposed method, we only insert adapters into the encoder side as we would like to modify the encoding function of y. It aims to encode the y into the same feature space as the semantically identical x. We also compare our choice to the common practice (Bapna and Firat, 2019b; Guo et al., 2020), where

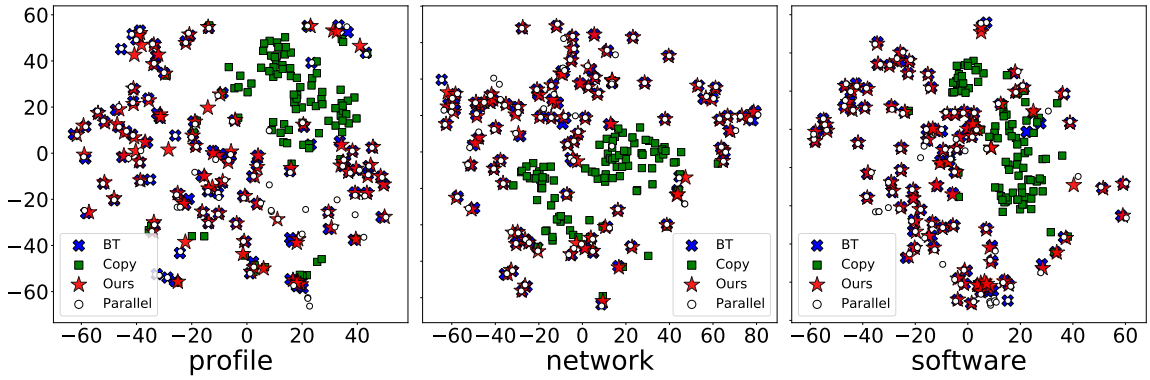

Figure 2: The t-SNE visualization (van der Maaten and Hinton, 2008) of the representation distributions for label *profile*, *network*, *software* in the datastores, which are created by different methods.

the adapters are inserted into both encoder and decoder sides. The results are shown in Table 4. We can observe that the adapters in the decoder side can only play a very limited role, which also demonstrates the motivation of our approach.

Adapter	IT	Medical	Law	Koran	Avg
Encoder	41.57	46.64	52.02	19.42	39.91
Encoder + Decoder	41.73	46.75	52.15	19.31	39.99

Table 4: The BLEU scores [%] of inserting adapters into encoder / encoder and decoder sides of our method.

Comparison with Fine-tuning Strategy. We compare our method with BT-FT, where the back-translation data is used for fine-tuning the full NMT model. The fine-tuning method easily causes catastrophic forgetting problem (Thompson et al., 2019) and results in performance degradation, especially when the data contains noise, as the results shown in Table 5.

Method	IT	Medical	Law	Koran	Avg
Basic NMT	38.35	39.99	45.48	16.26	35.02
UDA-kNN	41.57	46.64	52.02	19.42	39.91
BT-kNN	41.35	47.02	52.91	19.58	40.23
BT-FT	39.72	46.44	51.06	17.45	38.67

Table 5: The BLEU scores [%] of the non-parametric methods and fine-tuning method.

5 Conclusion

In this paper, we present UDA-kNN, a simple yet effective framework that directly utilizes monolingual data to construct in-domain datastore for unsupervised domain adaptation of kNN-MT. Experimental results verify that our method obtains significant improvement with target-side monolingual data. Our approach also achieves comparable performance with the BT-based method, while saving the extra cost of generating back-translation.

6 Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. This work is supported by National Science Foundation of China (No. U1836221, 61772261), National Key R&D Program of China (No. 2019QY1806) and Alibaba Group through Alibaba Innovative Research Program. We appreciate Weizhi Wang, Hao-Ran Wei, Yichao Du for the fruitful discussions. The work was done when the first author was an intern at Alibaba Group.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ankur Bapna and Orhan Firat. 2019a. Non-parametric adaptation for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1921–1931, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019b. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1417–1422, Hong Kong, China. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2018. Search engine guided neural machine translation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5133–5140. AAAI Press.

Junliang Guo, Zhirui Zhang, Linli Xu, Boxing Chen, and Enhong Chen. 2021. Adaptive adapters: An efficient way to incorporate BERT into neural machine translation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:1740–1751.

Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. 2020. Incorporating BERT into parallel sequence decoding with adapters. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest neighbor machine translation. *CoRR*, abs/2010.00710.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Hao-Ran Wei, Zhirui Zhang, Boxing Chen, and Weihua Luo. 2020. Iterative domain-repaired back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5884–5893, Online. Association for Computational Linguistics.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018a. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018b. Joint training for neural machine translation models with monolingual data. *CoRR*, abs/1803.00353.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. Adaptive nearest neighbor machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 368–374, Online. Association for Computational Linguistics.

A Appendix

A.1 Translation Examples

Table 6 shows a translation example selected from the Medical dataset. We can observe that our proposed UDA-kNN can make more proper word selections compared with basic NMT as well as Copy-kNN, thanks to the effective in-domain datastore construction. In addition, the overall translation accuracy of our method is close to BT-kNN and parallel-kNN, which utilize bilingual pairs to create datatsore while we only use monolingual pairs.

Source	Insbesondere bei großen chirurgischen Eingriffen ist eine genaue Überwachung der Substitutionstherapie mit Hilfe einer Koagulationsanalyse (Faktor VIII-Aktivität im Plasma) unbedingt erforderlich.
Reference	In the case of major surgical interventions in particular, precise monitoring of the substitution therapy by means of coagulation analysis (plasma factor VIII activity) is indispensable.
Basic NMT	Particularly in the case of major surgical procedures, precise monitoring of the substitution therapy with the help of a coagulation analysis (factor VIII activity in the plasma) is absolutely necessary.
Copy-kNN	Particularly in case of major surgical ***interventions*** a precise monitoring of the substitution therapy with the help of a coagulation analysis (factor VIII activity in the plasma) is necessary.
BT-kNN	In the case of major surgical ***interventions*** in particular, precise monitoring of the substitution therapy ***by means of*** a coagulation analysis (factor VIII activity in plasma) is ***indispensable***.
UDA-kNN	In the case of major surgical ***interventions*** in particular, precise monitoring of the substitution therapy ***by means of*** coagulation analysis (factor VIII activity in the plasma) is ***indispensable***.
Parallel-kNN	In the case of major surgical ***interventions*** in particular, precise monitoring of the substitution therapy ***by means of*** a coagulation analysis (plasma factor VIII activity) is ***indispensable***.

Table 6: Translation examples of different systems in Medical domain.