Domain Adaptation for NMT via Filtered Iterative Back-Translation

Surabhi Kumari, Nikhil Jaiswal, Mayur Patidar, Manasi Patwardhan,
Shirish Karande, Puneet Agarwal[1], Lovekesh Vig
TCS Research, New Delhi, India
{surabhi.kumari6, nikhil.jais, patidar.mayur,
manasi.patwardhan, shirish.karande, lovekesh.vig} @tcs.com
[1]dr.puneet.a@outlook.com

Abstract

Domain-specific Neural Machine Translation (NMT) model can provide improved performance, however, it is difficult to always access a domain-specific parallel corpus. Iterative Back-Translation can be used for fine-tuning an NMT model for a domain even if only a monolingual domain corpus is available. The quality of synthetic parallel corpora in terms of closeness to in-domain sentences can play an important role in the performance of the translation model. Recent works have shown that filtering at different stages of the back translation and weighting the sentences can provide state-of-the-art performance. In comparison, in this work, we observe that a simpler filtering approach based on a domain classifier, applied only to the pseudo-training data can consistently perform better, providing performance gains of 1.40, 1.82 and 0.76 in terms of BLEU score for Medical, Law and IT in one direction, and 1.28, 1.60 and 1.60 in the other direction in low resource scenario over competitive baselines. In the high resource scenario, our approach is at par with competitive baselines.

1 Introduction

Neural Machine Translation (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017) systems heavily rely on the availability of the parallel corpora to produce good quality translations (Koehn and Knowles, 2017). Even for high resource language pairs, in-domain parallel corpora are scarce. Chu and Wang (2018); Chu et al. (2018) address this challenge of domain adaptation with the objective of improving the performance of an NMT system by exploiting the in-domain monolingual corpora and out-of-domain parallel corpora for a given language pair.

In the current work, we build on top of the existing data centric approaches for domain adaptation (Chu and Wang, 2018), i.e., Back-Translation (BT) (Sennrich et al., 2016a) and Iterative Back-Translation (IBT) (Hoang et al., 2018). IBT is a variant of BT, which leverages both source and target-side monolingual corpora along with the out-of-domain parallel corpora and train $NMT_{s \to t}$ and $NMT_{t \to s}$ in alternate fashion till convergence, where $NMT_{s \to t}$ generates the synthetic parallel corpora for $NMT_{t \to s}$ and vice versa.

The performance of the NMT is influenced by the quality of synthetic parallel corpora as noted by Poncelas et al. (2018); Fadaee and Monz (2018). Hence, for the domain adaptation task, Dou et al. (2020) proposed a curriculum-based approach (*DDSWIBT*) for sentence selection from the in-domain monolingual corpora and use Junczys-Dowmunt (2018) for weight assignment to synthetic parallel corpora. In initial iterations of *IBT*, *DDSWIBT* prefer simple sentences over representative in-domain sentences, in later iterations they use more representative sentences as compared to simple sentences. Meanwhile, Imankulova et al. (2017) use an in-domain language model (*sent-LM*) and "Round-Trip BLEU" score for synthetic parallel corpora filtering.

In this paper, we propose a "classifier augmented filtered iterative back-translation" (*CFIBT*) for the domain adaptation task. We train two Convolutional Neural Network (CNN) (Kim, 2014) based binary classifiers, one in source and the other in target language on the combination of in-domain and out-of-domain corpora. We use *IBT* for synthetic parallel corpora generation and classifier-based filtering to remove the pair of sentences where the synthetic sentence in the pair does not belong to the domain. This entire procedure is depicted in Figure 1. We do not employ sentence selection over the monolingual corpora, or a weighting mechanism for the synthetic corpora, and neither utilize any "Round-Trip" criteria for scoring the synthetic parallel corpora.
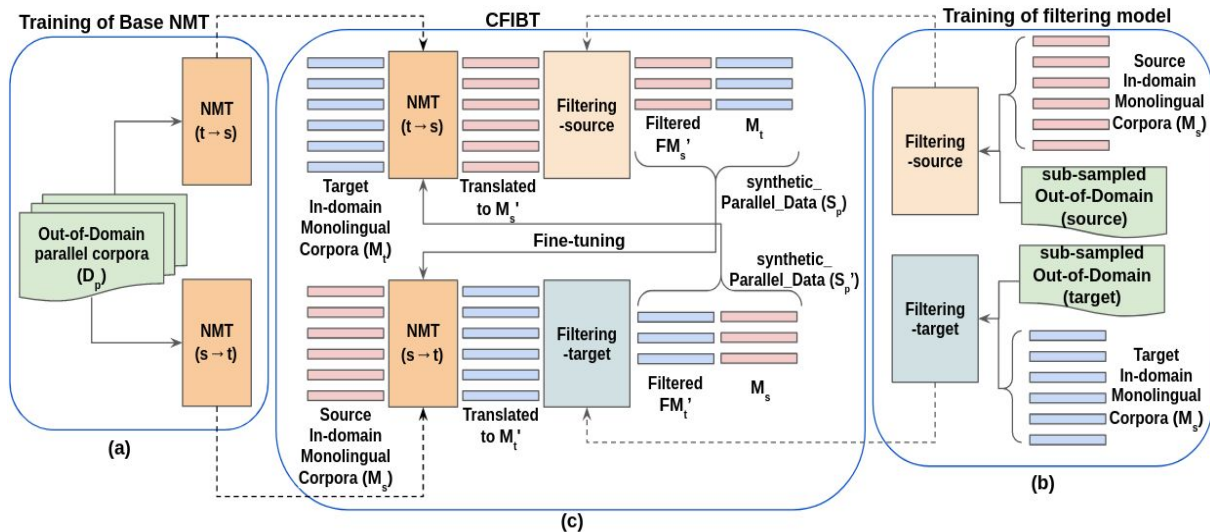
Figure 1: Overall procedure of our proposed approach. (a) We first train two Base NMT models - one in each direction on the out-of-domain parallel corpora. (b) We train two classifier-based filtering models - one for each source and target language to distinguish between in-domain and out-domain translated sentences. (c) We then use the trained NMT models to translate in-domain monolingual corpora. The translated sentences are then filtered to remove out of domain sentences. The remaining sentences along with their corresponding true source/target sentences are used to curate synthetic parallel data which is then utilized to fine-tune the NMT models and this entire cycle is iterated until convergence.

In the current work, we present our domain adaptation results for the German (de) - English (en) language pair on three different domains - Medical, Law and IT under low and high resource scenarios. In the low resource scenario, our proposed method *CFIBT* outperforms all the baselines in every domain. In the high resource settings, *CFIBT* outperforms the baselines in most of the scenarios, whereas it performs competitively with the best baseline results in the rest of the scenarios.

The rest of the paper is organized as follows: We describe related work in Section 2, our problem statement in Section 3, the proposed approach in Section 4. We present the results of the proposed and other baseline approaches in Section 5 and conclude in Section 6.

2 Related Work

In the current work, our objective is to improve the performance of the NMT model on in-domain sentences given out-of-domain parallel corpora and in-domain monolingual corpora in both source and target language, which is known as domain-adaptation for NMT (Chu and Wang, 2018; Chu et al., 2018). Approaches for domain-adaptation (Chu and Wang, 2018) are categorized into data-centric and model-centric. Data-Centric approaches for domain adaptation focuses on the use of in-domain monolingual corpora (Zhang and Zong, 2016; Cheng et al., 2016), synthetic corpora (Sennrich et al., 2016a;

Hoang et al., 2018; Hu et al., 2019), or parallel corpora (Luong and Manning, 2015; Chu et al., 2017) along with the out-of-domain parallel corpora. On the other hand, model-centric approaches modify the NMT architecture to include domain information i.e. domain-tags (Britz et al., 2017), domain embedding with word embeddings (Kobus et al., 2017) and assign higher weights to in-domain sentences as compared to out-of-domain sentences (Wang et al., 2017). In our current work, we use the data-centric approach for domain adaptation to generate synthetic parallel data via Iterative Back-Translation.

Shimodaira (2000); Jiang and Zhai (2007); Foster et al. (2010); Søgaard (2011); Sgaard (2013) have proposed different instance weighting based approaches for domain adaptation in NLP, where in-domain instances are assigned more weight as compared to the out-of-domain instances. In NMT, Poncelas et al. (2018); Fadaee and Monz (2018) observe that noisy sentences in synthetic parallel corpora can affect the performance of the translation model. Round-Trip BLEU score (Papineni et al., 2002) between the authentic and synthetic versions of the same sentence is used by Imankulova et al. (2017, 2019) to filter noisy synthetic corpora. And they also use the language model trained on in-domain monolingual data to filter the noisy sentences. Similarly, Jaiswal et al. (2020) use a semantic similarity technique based on the sentence

Dataset	Train	Dev	Test
High	4.5M	3K	3K
Low	100K	3K	3K

Table 1: Out-of-domain Dataset description in high and low resource scenario, where Train, Dev and Test set consists of bilingual sentences.

Dataset	Monolingual	Dev	Test
Medical	400K	2K	2K
Law	500K	2K	2K
IT	240K	2.5K	1.8K

Table 2: In-domain Dataset description for three domains Medical, Law and IT where Monolingual refers to in-domain sentences both in source and target language. Dev and Test set consists of in-domain bilingual sentences.

embeddings of the source and synthetic corpus to filter out noisy pairs. Instead of filtering the noisy sentences from the training data, He et al. (2016); Zhang et al. (2018); Wang et al. (2019) assign the lower weight to them during model training. In Dou et al. (2020), they use a variant of Moore and Lewis (2010) for data selection from in-domain monolingual corpora and use Junczys-Dowmunt (2018) for weight assignment to synthetic corpora generated by IBT. In the current approach, we use the whole in-domain monolingual data and filter the noisy synthetic corpora with the help of a simple binary classifier which is trained on in-domain and out-of-domain corpora.

3 Problem Description

Given out-of-domain parallel corpora D_p, and in-domain monolingual corpora M_s, M_t in source and target language respectively. Our objective is to create a pair of in-domain NMT models $NMT_{s \rightarrow t}$, $NMT_{t \rightarrow s}$ which can translate in-domain sentences with high efficacy from source to target ($s \rightarrow t$) and target to source language ($t \rightarrow s$) respectively. Similar to Edunov et al. (2018), we use more D_p in high resource scenario as compared to low resource scenario and the same amount of M_s, M_t in both the scenario.

4 Proposed Approach

We describe our proposed approach "classifier augmented filtered iterative back-translation" (*CFIBT*) in Algorithm 1. We assume that we have access to out-of-domain parallel corpora and in-domain

Algorithm 1 Filtering Augmented Iterative Back-Translation for Domain Adaptation

Require: D_p ▷ Out-of-Domain parallel Corpora
Require: M_s ▷ In-Domain monolingual Corpora in source language
Require: M_t ▷ In-Domain monolingual Corpora in target language
1: $NMT_{s \rightarrow t} \leftarrow$ Train NMT using D_p for $s \rightarrow t$
2: $NMT_{t \rightarrow s} \leftarrow$ Train NMT using D_p for $t \rightarrow s$
3: **while** $NMT_{s \rightarrow t}$ and $NMT_{t \rightarrow s}$ not converged **do**
4:      $M'_s \leftarrow$ Translate M_t using $NMT_{t \rightarrow s}$
5:      $M'_t \leftarrow$ Translate M_s using $NMT_{s \rightarrow t}$
6:      $FM'_s \leftarrow$ Filtering(M'_s)
7:      $FM'_t \leftarrow$ Filtering(M'_t)
8:      $S_p \leftarrow$ synthetic_Parallel_Data(FM'_s, M_t)
9:      $S'_p \leftarrow$ synthetic_Parallel_Data(FM'_t, M_s)
10:      $NMT_{s \rightarrow t} \leftarrow$ Fine-tune $NMT_{s \rightarrow t}$ using S_p
11:      $NMT_{t \rightarrow s} \leftarrow$ Fine-tune $NMT_{t \rightarrow s}$ using S'_p
12: **end while**
13: **return** $NMT_{s \rightarrow t}$, $NMT_{t \rightarrow s}$

monolingual corpora in both source and target languages. Firstly, we train $NMT_{s \rightarrow t}$ and $NMT_{t \rightarrow s}$ using out-of-domain parallel corpora D_p. Then we use these trained $NMT_{s \rightarrow t}$, $NMT_{t \rightarrow s}$ for translating M_s to M'_t and M_t to M'_s respectively. We then apply a classifier-based filtering technique which is described below on these translated sentences. The filtered sentences FM'_s and FM'_t and their corresponding in-domain monolingual sentences M_t and M_s are used to curate synthetic parallel data. Thereafter, $NMT_{s \rightarrow t}$ and $NMT_{t \rightarrow s}$ are fine-tuned on this synthetic parallel data. This entire process repeats until convergence. We consider $NMT_{s \rightarrow t}$ and $NMT_{t \rightarrow s}$ to be converged when there is no improvement in both the models when compared to its preceding iteration model.

Filtering Model: We propose a naive "classifier-based filtering model" using a Convolutional Neural Network. The filtering model consists of a binary classifier trained to distinguish between in-domain and out-of-domain sentences. For each domain, we train two such models - one for source and other for the target language. The classifier is trained only once at the beginning, and we utilize the same classifier in each successive iterations. Given a translated sentence as input, the classifier

Source (de)	warnhinweis, dass das arzneimittel fr kinder unerreichbar und nicht sichtbar aufzubewahren ist
Target (en)	special warning that the medicinal product must be stored out of the reach and sight of children
BASE	warning that drugs for children is unvisible and not visible .
CFIBT$_1$	warning that the medicinal product is being unabsorbed and has not been visible .
CFIBT$_2$	warning that the medicines for children are not being able and not visible .
CFIBT$_3$	warning that the medicinal product must be stored out of the reach and sight of children

Table 3: Translation of a German (de) sentence to English (en) via different *CFIBT*$_i$ models, where i represent the ith iteration of *CFIBT*.

predicts the probability of this sentence as being in-domain. All translated sentences having a probability greater than a certain threshold are considered in-domain sentences.

5 Results and Discussion

Here, we describe the datasets and the training details of our experiments. We also discuss and analyze the results and key observations.

5.1 Dataset Description

We perform our experiments on German-English (de-en) language pair. In both low and high resource scenarios, we use the same out-of-domain News dataset as used by Dou et al. (2020), which is described in Table 1.

For in-domain data, as described in Table 2, we use the same dev and test set as used by Dou et al. (2020) for Medical (EMEA), Law (Acquis) and also the same number of monolingual sentences during domain adaptation. In addition to Medical and Law, in the current work, we also report results on the IT domain and for that, we use the dataset described in Tiedemann (2012).

We tokenize the out-of-domain sentence pairs as well as in-domain sentences using moses (Koehn et al., 2007) and apply byte-pair-encoding (Sennrich et al., 2016b) with $37K$ merge operations.

5.2 Training Details

We train two types of models i.e. filtering models and NMTs.

Filtering Models: We train the filtering models in English and German languages for each domain. We use language model and classifier as filtering models for *sent-LM* and *CFIBT* respectively. For the *language model*, we use one layer LSTM (Hochreiter and Schmidhuber, 1997) having 512 embedding size and 50 sequence length using TF-LM Toolkit (Verwimp et al., 2018). The model is trained till convergence with the patience of three. We train on the tokenized monolingual in-domain dataset for each domain in English and German with vocabulary size $\approx 60K$ and $\approx 80K$ respectively. The *Classifier* architecture is inspired by Kim (2014). For training the binary classifier, we use sub-sampled out-of-domain data as one class and in-domain as another. The tokenized sentences are used with a vocabulary size of $50K$. For filtering models, we obtain the optimal values of thresholds based on the development set, where the objective is to maximize the true positives (i.e. in-domain sentences) and minimize the false positives (i.e. out-of-domain sentences) in the synthetic parallel corpus. The overall intuition is that the classifier should help to select the in-domain sentences, which could then be utilized to further train the NMT models. In *sent-LM* for all domains, we use 60 and 80 as a value of perplexity to filter-out out-of-domain sentences from synthetic sentences in English and German respectively. For *CFIBT*, we use 0.6 for Medical and 0.5 for Law and IT as a threshold over classifier probability for filtering sentences in English and German.

NMT: We use Base-Transformer (Vaswani et al., 2017) for our experiments. We use FairSeq (Ott et al., 2019) for training all the NMT models. We use out-of-domain parallel corpora to train the initial NMT model i.e. *BASE* in both low and high resource scenarios. We finetune the model obtained from *BASE* in *BT*, *IBT*, *sent-LM* and *CFIBT* with the synthetic parallel in-domain dataset, curated with respective approaches. We use the value of patience as five for all approaches.

5.3 Results and Analysis

Here, we compare and discuss the results of our proposed approach along with other baseline methods. As shown in Table 4, we compare the BLEU scores in two different scenarios,

Domain →	Medical		Law		IT	
LP →	de-en	en-de	de-en	en-de	de-en	en-de
High Resource						
BASE	33.61	24.98	33.07	23.33	21.93	16.27
BT	41.05	36.32	38.27	28.32	35.31	24.80
sent-LM	47.44	37.85	40.82	30.35	39.24	**30.11**
IBT	**47.71**	**38.01**	39.46	29.04	38.93	29.37
DDSWIBT	45.46	36.45	39.11	29.04	-	-
CFIBT	47.59	37.61	**40.99**	**30.38**	**40.06**	29.93
Low Resource						
BASE	10.05	6.53	8.52	6.84	5.05	3.70
BT	22.64	14.02	18.47	10.53	13.07	10.79
sent-LM	36.21	29.35	25.36	17.28	28.80	24.32
IBT	33.14	24.31	22.96	14.31	28.06	24.16
DDSWIBT	31.22	28.12	22.06	13.28	-	-
CFIBT	**37.61**	**30.63**	**27.18**	**18.88**	**29.56**	**25.82**

Table 4: We use the BLEU (Papineni et al., 2002) score to compare *CFIBT* with existing baselines, *BASE* (Vaswani et al., 2017), *BT* (Sennrich et al., 2016a), *sent-LM* (Imankulova et al., 2017), *IBT* (Hoang et al., 2018) and *DDSWIBT* (Dou et al., 2020). Except *DDSWIBT*, we implemented all baselines.

viz., high resource and low resource, on three different domains i.e. Medical, Law and IT in both directions for German-English (de-en) language pair on in-domain test set. With monolingual data only in both source and target language, we get performance gains of 27.56, 18.66 and 24.51 in terms of BLEU score for Medical, Law and IT in one direction (de-en), and 24.1, 12.04 and 22.12 in the other direction (en-de) in low resource scenario over the *BASE*. In the low resource scenario, *CFIBT* outperformed *sent-LM* in both directions and all the domains. *CFIBT* also outperformed *sent-LM* in high resource scenario except in one direction for the IT domain. In the low resource scenario, filtering based approaches performs better than the IBT. And *CFIBT* outperformed in all the cases. Our results show that *CFIBT* is efficient when the base model is not adequately trained. In high resource scenario results of *CFIBT* are comparable with other baselines. *CFIBT* outperformed in both directions for Law and one direction IT domain, respectively.

Why *CFIBT* works? According to Figure 2 (Appendix), IBT is trained with all synthetic bilingual sentences without filtering, which may hurt the performance of IBT in subsequent iterations because the current model is used to generate the data for the next iteration. But in *CFIBT*, as shown in Figure 3, filtering prevents

training of the NMT model on out-of-domain sentence pairs which leads to a better domain model in subsequent iterations.

6 Conclusion

In the context of domain adaptation for NMT, we propose a simple and effective approach for filtering the synthetic parallel corpus, which is as good as more involved approaches for the same task. In the low resource scenario, the proposed approach outperforms all the existing baselines whereas we get similar results to baselines in the high resource scenario. As a part of future work, we would like to validate our findings on different language pairs and multiple domains and also would like to explore the combination of different filtering techniques. And instead of training different filtering models for source and target language, we would like to use a single multilingual filtering model.

Acknowledgements

References

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.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark. Association for Computational Linguistics.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. 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 (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2018. A comprehensive empirical comparison of domain adaptation methods for neural machine translation. *Journal of Information Processing*, 26:529–538.

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*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5894–5904, Online. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 820–828. Curran Associates, Inc.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, pages 1735–80.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2019. Filtered pseudo-parallel corpus improves low-resource neural machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(2).

Nikhil Jaiswal, Mayur Patidar, Surabhi Kumari, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, and Lovekesh Vig. 2020. Improving NMT via filtered back translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 154–159, Suzhou, China. Association for Computational Linguistics.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the*

2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. IN-COMA Ltd.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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, Vancouver. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Alberto Poncelas, Dimitar Sht. Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *CoRR*, abs/1804.06189.

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. Association for Computational Linguistics.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 682–686, Portland, Oregon, USA. Association for Computational Linguistics.

Anders Sgaard. 2013. Semi-supervised learning and domain adaptation in natural language processing. *Synthesis Lectures on Human Language Technologies*, 6(2):1–103.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Lyan Verwimp, Hugo Van hamme, and Patrick Wambacq. 2018. TF-LM: TensorFlow-based language modeling toolkit. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. 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 791–802, Hong Kong, China. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Zhirui Zhang, S. Liu, Mu Li, M. Zhou, and E. Chen. 2018. Joint training for neural machine translation models with monolingual data. In *AAAI*.

A   Appendix

A.1   Results and Analysis

Figure 2 describes the variation in the number of synthetic parallel sentence pairs as well as BLEU score on test data with the different number of iterations for the *IBT*, *sent-LM* and *CFIBT* models for low resource scenario. For *IBT*, the number of synthetic sentence pairs remains the same for all the iteration. For filtering models, the number of filtered synthetic sentences increases with every upcoming iteration. This increase in good quality synthetic pairs helps in the improvement of the translation model resulting in a better BLEU score. This process is iterated until we do not observe any improvement in the BLEU score compared to the last iteration. Since one direction translation model creates synthetic data for the other direction, we observe that if there is an improvement in the former case for a given iteration, then there is an improvement in the later model for the next iteration.
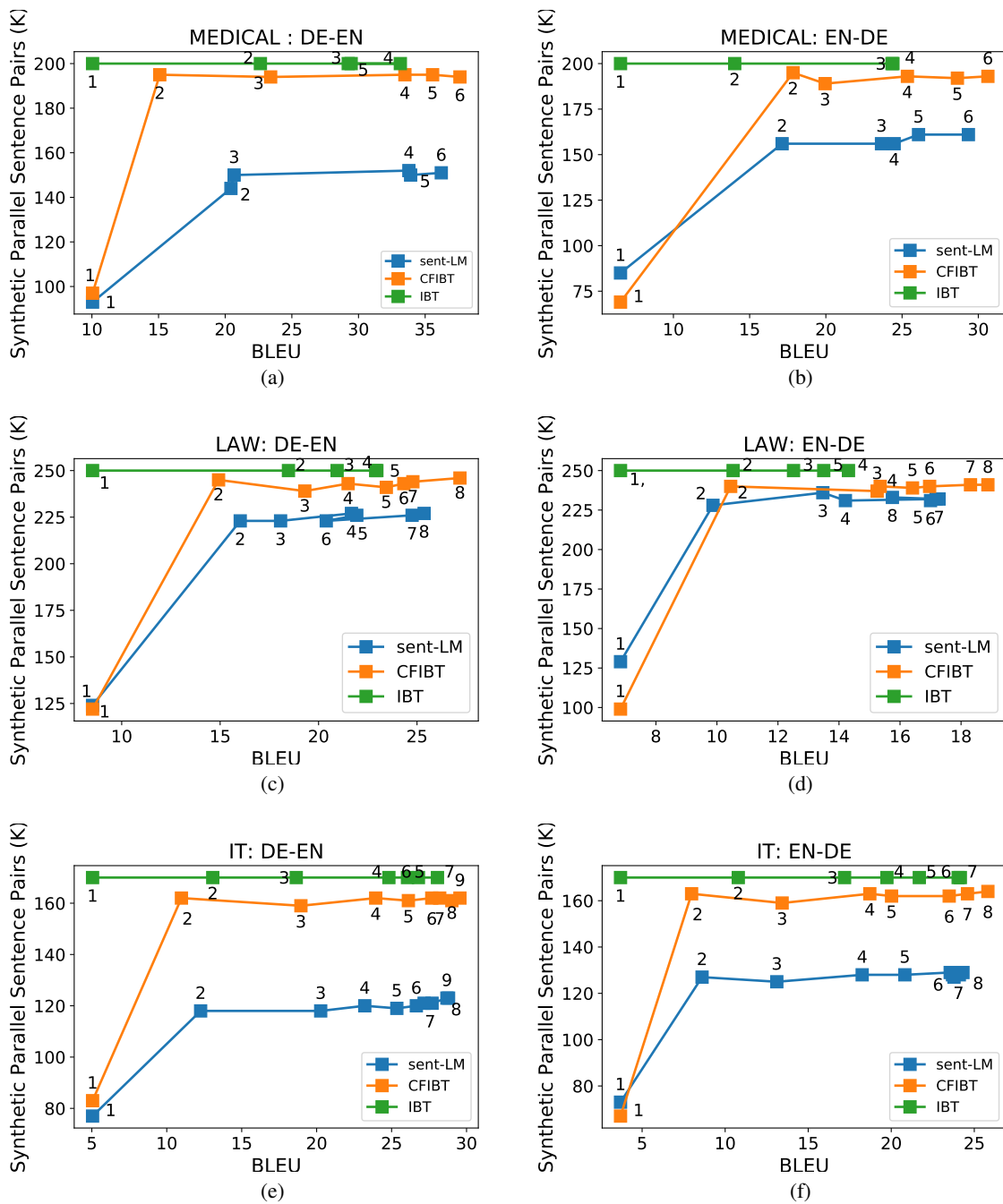
Figure 2: We represent the variation in Synthetic Parallel Sentence Pairs and BLEU scores for different number of iterations in *IBT*, *CFIBT* and *sent-LM* models for low resources. The numerical value represents the iteration number for that particular model (a) & (b) represents de-en and en-de models for Law domain, (c) & (d) represents de-en and en-de models for Medical domain, (e) & (f) represents de-en and en-de models for IT domain