Optum's Submission to WMT22 Biomedical translation tasks

Sahil Manchanda
sahil_manchanda@optum.com

Saurabh Bhagwat
saurabh_bhagwat@optum.com

Abstract

This paper describes Optum's submission to the Biomedical Translation task of the seventh conference on Machine Translation (WMT22). The task aims at promoting the development and evaluation of machine translation systems in their ability to handle challenging domain-specific biomedical data. We made submissions to two sub-tracks of ClinSpEn 2022, namely, ClinSpEn-CC (clinical cases) and ClinSpEn-OC (ontology concepts). These sub-tasks aim to test translation from English to Spanish. Our approach involves fine-tuning a pre-trained transformer model using in-house clinical domain data and the biomedical data provided by WMT. The fine-tuned model results in a test BLEU score of 38.12 in the ClinSpEn-CC (clinical cases) subtask, which is a gain of 1.23 BLEU compared to the pre-trained model.

1 Introduction

The quality of Neural Machine Translation (NMT) was boosted by the use of Recurrent Neural Networks (RNN) for machine translation. In this approach, the source sentence is fed to an encoder which outputs a context vector. This context vector is fed to the decoder to output the target language text (Cho et al., 2014). Some approaches also use Long Short Term Memory (Hochreiter and Schmidhuber, 1997) for this task (Sutskever et al., 2014).

Machine Translation (MT) systems after seeing great progress in recent years have been found to be sensitive to synthetic and natural noise in input, distributional shift, and adversarial examples (Koehn and Knowles, 2017; Belinkov and Bisk, 2017; Durrani et al., 2019; Anastasopoulos et al., 2019; Michel et al., 2019). Fine-tuning has proven to be a successful technique to carry out this task. One of the most prominent variations is described in (Chu and Wang, 2018), which trains an NMT model on out-of-domain corpora until model convergence and then resumes training from step 1 on a mix of in-domain and out-of-domain data.

A fine-grained human evaluation research of the transformer based systems and state-of-the-art recurrent systems was carried out on the translation from English to Chinese. The evalution results shows reduction in errors by 31 percent and significantly less errors in 10 out of 22 error categories when using Transformer based MT systems. (Ye and Toral, 2020). Another research has shown that improved efficiency and accuracy can be obtained by converting a pre-trained transformer into its efficient recurrent counterpart. A swap procedure is implemented which replaces softmax attention of a pertained transformer with its linear-complexity recurrent alternative followed by fine-tuning. Fine-tuning has proven to help reduce the training cost and improve efficiency and accuracy (Kasai et al., 2021).

We took part in WMT 2022 Biomedical translation task from English to Spanish using the fine-tuning approach on the Transformer based models and we describe our efforts in this paper. The paper is structured as follows. The data sets and their preparation is outlined in Section 3 and Section 4, followed by details of the experiments carried out and their results in Section 5. We then present the summary of our findings and conclusion in Section 6.

2 Related Work

Machine translation systems out of domain performance has been negatively impacted to the extent that they completely sacrifice adequacy for the sake of fluency. Hence, the presence of domain inconsistency is considered a key challenge in machine translation (Koehn and Knowles, 2017). The common approach to tackle this challenge is firstly to train an MT system on a (generic) source domain and secondly to fine-tune it on a (specific) target domain (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Servan et al., 2016; Chu et al., 2017), followed by continuous fine-tuning of data

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 925 - 929

sets which are similar to the target domain (Sajjad et al., 2017), or to dynamically change the balance of data towards the target domain (van der Wees et al., 2017). An alternative approach is to train systems on multiple domains simultaneously, while adding domain-specific tags to the input examples (Kobus et al., 2016).

Other methods include the works around Dual Contextual (DC) module, which is an extension of the conventional self-attention unit, to effectively make use of both, local and global contextual information. This work aims to further improve the sentence representation ability of the encoder and decoder sub-networks, thus enhancing the overall performance of the translation model (Ampomah et al., 2021). Domain adaptation methods include instance weighing, data selection (Wang et al., 2017) and incorporating a domain classifier (Chen et al., 2017; Britz et al., 2017).

Some language pairs do not have enough parallel text for training. Hence, to counter the data sparsity problem of the NMT training some have used various strategies like augmenting training data, exploiting training data from other languages, alternative learning strategies that use only monolingual data (Haque et al., 2021). Some of the researchers have made use of monolingual data available either in the target domain, for example, by training the decoder on these data sets (Domhan and Hieber, 2017), or by back-translating (Sennrich et al., 2016), or in the source domain, using similar techniques (Zhang and Zong, 2016).

3 Data

In the experiments described in this paper, we use data sets from both the general and clinical domains. ParaCrawl, EMEA, and WMT are available in the public domain, while, M&R Letters is a data set internal to Optum. The M&R in-domain data set comprises of medical claim correspondence letters sent to the insurance customers which have been manually translated to Spanish. Among the public data sets, ParaCrawl is the largest publicly available parallel corpora for European languages. EMEA is a multi-lingual parallel corpus made out of PDF documents from the European Medicines Agency. We have used data from all three subtracks namely, clinical cases, clinical terminology, and ontology concepts of the ClinSpEn data set provided by WMT. Table 1 summarizes the data sets used and their size. It is important to note that we generate train and test splits on ParaCrawl (general domain), EMEA, and M&R data sets (clinical domain) and evaluate on these. For WMT, we use all 8K sentence pairs as training data and share evaluation BLEU scores computed by WMT submission system on their hidden test set.

Data Fragment	Sentences	Domain
ParaCrawl	38M	General
M&R Letters	492K	Medical
EMEA	15K	Clinical
WMT	8K	Clinical

Table 1: Data sets used in this work and corresponding source and number of sentences in each.

4 Data Preparation

The Data preparation very closely follows the steps outlined in (Manchanda and Grunin, 2020). The additional steps are listed below.

1. **Language Check Elimination**
   Sentences not from the intended language were eliminated.

2. **Length difference check**
   The internal data that we used comprised of correspondence letters to our customers anonymized and their manual translations. It was found upon a close observation that manual translations differ depending on the translator. Sometimes, the same phrase can be translated multiple ways or some additional information can be added unintentionally to the translation which can confuse the learning algorithm leading to under-fitting. We eliminated any translation that differs from the source sentence in length by more than 40 percent.

5 Experiments and Results

As described in the Data Section (3), We are using data sets from both the General domain and Medical/Clinical domain. To fine-tune the model, we have a 2-GPU setup with a docker container deployed on on-premise machines containing all the required packages to fine-tune the OPUS en-es translation model [1]. We use HuggingFace transformers library (Wolf et al., 2020) for all our experiments.

[1] https://huggingface.co/Helsinki-NLP/opus-mt-en-es

The following fine-tuning experiments are done on the transformer model used by the Helsinki-NLP opus-mt-en-es model. As evident from its model card, this model was trained on general-purpose English to Spanish training corpus and in these experiments, we will try to fine-tune the model to the clinical domain.

Since the data provided by the sub-task was limited, we used the entire WMT 2022 data as training data and used train-test splits on other clinical domain data sets to test the success of fine-tuning.

One of our key observation while doing the experiments and serving these models on production systems were that they regularly need to be checked for over-fitting and hallucination errors. In addition to evaluation by BLEU scores, We do a "**Sanity check**" by running an inference with source language strings of various lengths to mimic handwritten text and check if the translation is not adding extra tokens.

1. **Experiment 0: Reference Baseline**
   We use the model already pre-trained without any fine-tuning as our reference baseline and compare our fine-tuning results against this to determine the better models.

2. **Experiment 1: Mix of General and In-domain data**
   First, we fine-tune the general purpose model on a mix of in-domain and public data set. Our in-domain data sets are M&R correspondence letters, EMEA clinical data set and WMT 2022. We mix these with 2 million sentences randomly selected from the ParaCrawl corpus to keep the model from over-fitting to only one domain. We keep the learning rate on the higher side (1e-5) for this experiment and train for 1 epoch only. We do not add length difference check (2) in this experiment on the in-domain data.

3. **Experiment 2: Fine-tuning on only In-domain data**
   Our next experiment was to fine-tune the public model on only the in-domain data sets. This experiment contains all the data preparation steps. The learning rate for this experiment was kept lower as compared to the previous experiment (1e-6) as the data was purely in-domain.

Figure 1 shows a graph of the BLEU scores at evaluation time for all the above-mentioned experiments. Along with the BLEU scores on the test splits of general and Clinical (EMEA/M&R) datasets, this figure also shows the test BLEU scores provided by WMT on their hidden test sets. We observe that the model trained on only general-purpose data (Experiment 0) performs decently on both in-domain and general-purpose data sets. Experiments 1 and 2 yield better results on the EMEA/M&R data sets, and degrade a little on the general-purpose data sets. It can be noted that both experiments have the same scores on general and EMEA/M&R datasets. This indicates that the approach of fine-tuning with a high learning rate with some general domain data present (experiment 1) and fine-tuning with a low learning rate only on the in-domain data (experiment 2) yields very similar results.

However, Experiment 2 yields the best results on the WMT test data set and hence is our primary submission to the task. It is interesting to note that the gain on the BLEU scores of EMEA and M&R datasets is more significant as compared to the gain in WMT BLEU scores. One of the major reasons for that could be the amount of data available for this particular domain.

6 Conclusion

We fine-tuned a publicly available model in multiple ways using different combinations of data from various sources. We showed how fine-tuning is sensitive to new domains and can show promising results if done diligently. This paper shows the results of fine-tuning on a single domain but we think that fine-tuning on any new domain would provide gains in the translation quality. The scale of this gain, however, can depend on the amount of training data available in that particular domain.

7 Limitation

As evident from our experiments and results, in-domain machine translation involves some trade-off in translation quality amongst domains. When we tried to fine-tune a translation model to a new domain, the BLEU scores on the general domain drop. The users of the fine-tuned model need to be cognizant of the fact that while these models are the best for the domain they were fine-tuned for, they might not be the best to translate general handwritten text which lacks the structure of the fine-tuning data. We recommend separate specialized models for different use cases.

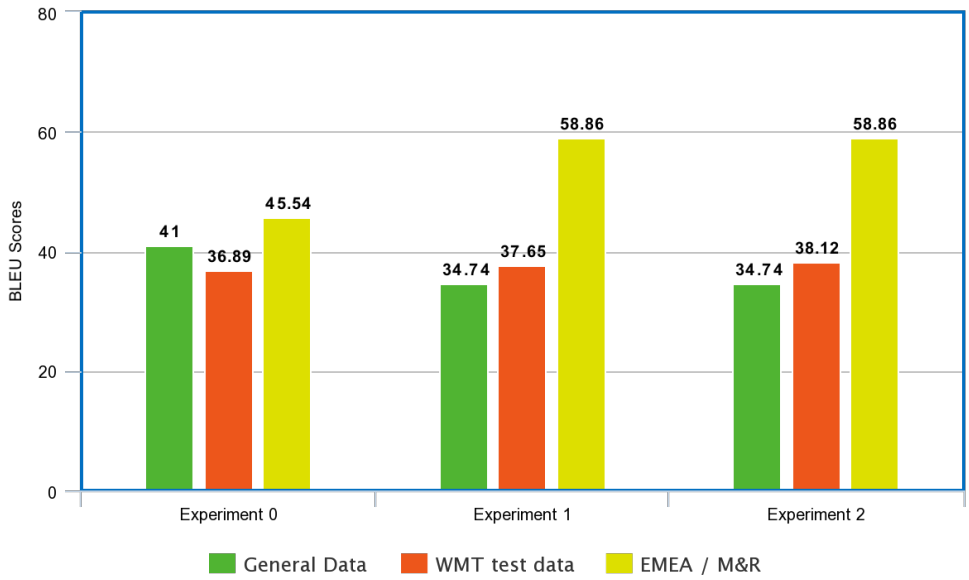

Figure 1: BLEU scores on evaluation data sets

References

Isaac Kojo Essel Ampomah, Sally McClean, and Glenn Hawe. 2021. Dual contextual module for neural machine translation. *Machine Translation*, 35(4):571–593.

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. 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 3070–3080, Minneapolis, Minnesota. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *CoRR*, abs/1711.02173.

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.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46, Vancouver. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *CoRR*, abs/1701.03214.

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.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. Association for Computational Linguistics.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. 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 1504–1516, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897.

Rejwanul Haque, Chao-Hong Liu, and Andy Way. 2021. Recent advances of low-resource neural machine translation. *Machine Translation*, 35.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Jungo Kasai, Hao Peng, Yizhe Zhang, Dani Yogatama, Gabriel Ilharco, Nikolaos Pappas, Yi Mao, Weizhu Chen, and Noah A. Smith. 2021. Finetuning pretrained transformers into rnns. *CoRR*, abs/2103.13076.

Catherine Kobus, Josep Maria Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *CoRR*, abs/1612.06140.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *CoRR*, abs/1706.03872.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Sahil Manchanda and Galina Grunin. 2020. Domain informed neural machine translation: Developing translation services for healthcare enterprise. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 255–261, Lisboa, Portugal. European Association for Machine Translation.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. 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 3103–3114, Minneapolis, Minnesota. Association for Computational Linguistics.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017. Neural machine translation training in a multi-domain scenario. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 66–73, Tokyo, Japan. International Workshop on Spoken Language Translation.

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, Berlin, Germany. Association for Computational Linguistics.

Christophe Servan, Josep Maria Crego, and Jean Senellart. 2016. Domain specialization: a post-training domain adaptation for neural machine translation. *CoRR*, abs/1612.06141.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuying Ye and Antonio Toral. 2020. Fine-grained human evaluation of transformer and recurrent approaches to neural machine translation for english-to-chinese. *CoRR*, abs/2006.08297.

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.