Multi-Domain Adaptation in Neural Machine Translation Through Multidimensional Tagging

Emmanouil Stergiadis emmanouil.stergiadis@booking.com
Satendra Kumar satendra.kumar@booking.com
Fedor Kovalev fedor.kovalev@booking.com
Pavel Levin pavel.levin@booking.com

Abstract

While NMT has achieved remarkable results in the last 5 years, production systems come with strict quality requirements in arbitrarily niche domains that are not always adequately covered by readily available parallel corpora. This is typically addressed by training domain specific models, using fine-tuning methods and some variation of back-translation on top of in-domain monolingual corpora. However, industrial practitioners can rarely afford to focus on a single domain. A far more typical scenario includes a set of closely related, yet succinctly different sub-domains. At Booking.com, we need to translate property descriptions, user reviews, as well as messages, (for example those sent between a customer and an agent or property manager). An editor might need to translate articles across a set of different topics. An e-commerce platform would typically need to translate both the description of each item and the user generated content related to them. To this end, we propose MDT: a novel method to simultaneously fine-tune on several sub-domains by passing multidimensional sentence-level information to the model during training and inference. We show that MDT achieves results competitive to N specialist models each fine-tuned on a single constituent domain, while effectively serving all N sub-domains, therefore cutting development and maintenance costs by the same factor. Besides BLEU (industry standard automatic evaluation metric known to only weakly correlate with human judgement) we also report rigorous human evaluation results for all models and sub-domains as well as specific examples that better contextualise the performance of each model in terms of adequacy and fluency. To facilitate further research, we plan to make the code available upon acceptance.

1 Introduction

Neural machine translation (NMT) has achieved remarkable results in recent years. A strong testament to its success and efficacy is the increasingly widespread industrial adoption of NMT solutions Johnson et al. (2017); Levin et al. (2017a); Crego et al. (2016). Model parameter estimation in NMT architectures (Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) is still largely a supervised learning problem which requires large amounts of translated sentence pairs (parallel data). Obviously, acquiring a sufficient number of high quality parallel sentences in order to train a functional domain-specific NMT system can be prohibitively expensive; especially, if one needs to develop such systems for several domains across different language pairs. On the other hand, large quantities of untranslated in-domain content (monolingual data) are often readily available.

Various domain adaptation strategies have been developed to address the low-resource setting of niche domains (Chu and Wang, 2018). Some of the more popular approaches involve

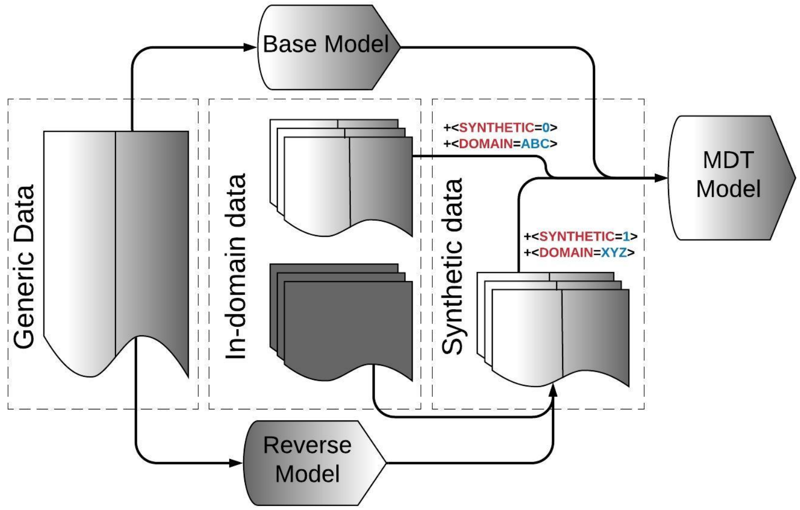

Figure 1: Schematic diagram of MDT in our setting. We use generic parallel data to train a base source-target and a reverse target-source models. We then back-translate target language monolingual in-domain data using the reverse model, and mix it with upsampled in-domain parallel data to fine-tune the base model. The data is tagged with two special tokens: <SYNTHETIC={0,1}>, and <DOMAIN={reviews,messaging,descriptions}>.

generating synthetic in-domain data with the help of existing monolingual corpora, and using that data to fine-tune the more general NMT systems Sennrich et al. (2016a).

In real-world scenarios practitioners often need to deploy translation engines for several closely related, yet different sub-domains. For example, an online travel marketplace needs to translate offering descriptions, user-generated reviews and customer service communications, all related to travel, but all having different linguistic nuances. This fragmentation is further compounded by the company's need to provide services across many distinct languages. It can be very expensive or outright impossible to develop and maintain separate translation pipelines for every combination of language and sub-domain.

We propose a new method for training models which are simultaneously fine-tuned on several closely related, yet succinctly different sub-domains. We show that those models achieve competitive (and often superior) results to single domain fine-tuned baselines while effectively serving N use cases, therefore cutting development and maintenance costs by a factor of N.

2 Related Work

Our work builds on a growing body of domain adaptation research, mainly related to fine-tuning through tagged back-translation.

2.1 Domain tagging

There are a number of research directions related to using tags (or special tokens) within NMT, primarily as a way to pass additional information to the model. Practically speaking, these are attractive approaches as they usually do not require any special modifications to off-the-shelf translation software. The majority of use cases tag sentences on the *source* side: Kobus et al. (2017) use them to control domain, Sennrich et al. (2016) the politeness, Yamagishi et al.

	Arabic	German	Russian
Parallel			
Generic	71M	92M	87M
Reviews	98k	63k	136k
Messaging	73k	76k	87k
Descriptions	60k	72k	80k
Monolingual			
Reviews	1M	1M	1M
Messaging	1M	1M	1M
Descriptions	1M	1M	1M

Table 1: Parallel and monolingual sentences used in our experiments.

(2016) the voice and Elaraby et al. (2018) the gender of translations. The idea also features in multilingual NMT models, for example Johnson et al. (2017) tag training examples according to which translation pair they belong to. An alternative approach by Britz et al. (2017) prepends the domain tag to the training input on the *target* side, thus forcing the decoder to predict the domain based on the source sentence alone.

2.2 Back-Translation for Domain Adaptation

Back-translation (BT) is a form of semi-supervised learning that can be used to fine-tune both statistical Bertoldi and Federico (2009); Bojar and Tamchyna (2011) and neural (Sennrich et al., 2016a) machine translation models to new domains. The idea behind this technique is to augment limited parallel in-domain data with a synthetic corpus produced by translating monolingual data from the *target* language using a *target-to-source* translation system. A synthetic corpus produced via back-translation will have machine-generated source sentences "translated to" human-written in-domain targets. BT model fine-tuning then becomes a three-stage process: first, genuine parallel data is used to train a reverse model in the target-to-source direction; second, that reverse model is used to translate target-side in-domain monolingual data into the source language; third, synthetic data is used in combination with few truly parallel in-domain samples to fine-tune the base source-to-target model. This simple approach works surprisingly well in practice Bojar et al. (2018); Barrault et al. (2019).

Recent research showed that the details of how we generate the synthetic BT data matter a lot (Edunov et al., 2018; Imamura et al., 2018). Specifically, the authors find that randomized sampling and noising is preferable to plain beam search. Edunov et al. (2018) hypothesise that the improvement is due to randomization contributing to the source-side diversification of the synthetic data. Caswell et al. (2019), on the other hand, suggest that synthetic data adds both helpful and harmful signals, which sampling and noising BT strategies help the model to separate. The TaggedBT technique which they introduce achieves competitive results by simply tagging synthetic data with a special token indicating that the data is machine-generated.

3 Multidimensional tagging

As discussed in Section 2, introducing special tokens in the training data has been independently useful at passing content-specific information (e.g. domain, voice, gender, etc.) and data-specific information (e.g. whether a given data point is synthetic). The current work extends this idea into the multidimensional setting. Whenever several meaningful dimensions describing the data are available at inference and training time, we can encode that information with special tokens indicating the values along each of the dimensions (Figure 1).

	Reviews			Messaging			Descriptions			Average		
	AR	DE	RU	AR	DE	RU	AR	DE	RU	AR	DE	RU
Human score												
Base model	3.65	3.73	3.50	3.27	3.44	3.18	2.67	3.28	2.95	3.20	3.48	3.21
+top10	**3.75 (+.10)**	3.80 (+.07)	3.57 (+.07)	3.36 (+.09)	3.65 (+.19)	3.53 (+.35)	3.02 (+.35)	3.70 (+.42)	2.95 (+.00)	3.38 (+.18)	3.71 (+.23)	3.47 (+.14)
+MDT	3.72 (+.07)	**3.88 (+.15)**	**3.62 (+.12)**	**3.49 (+.22)**	**3.78 (+.34)**	**3.53 (+.35)**	**3.20 (+.53)**	**3.73 (+.45)**	**3.04 (+.09)**	**3.47 (+.27)**	**3.80 (+.31)**	**3.40 (+.19)**
BLEU score												
Base model	42.95	43.63	38.25	39.01	44.18	41.18	45.00	45.97	38.92	42.32	44.60	39.45
+top10	**42.95 (+0.00)**	44.99 (+1.36)	38.35 (+0.10)	41.93 (+2.92)	**50.19 (+6.01)**	41.15 (-0.03)	45.35 (+0.35)	**50.98 (+5.01)**	37.84 (-1.08)	43.41 (+1.09)	48.72 (+4.13)	39.11 (-0.34)
+MDT	42.61 -0.34	**46.34 (+2.71)**	**41.12 (+2.87)**	**47.09 (+8.08)**	49.85 (+5.67)	**43.19 (+2.01)**	**46.54 (+1.54)**	50.84 (+4.87)	**39.14 (+0.22)**	**45.41 (+3.09)**	**49.01 (+4.41)**	**41.15 (+1.70)**

Table 2: Human evaluations and BLEU scores for the multi-domain adaptation experiments. MDT (our method) is competitive (and on average superior) against the strong fine-tuning baseline (*top10* from (Edunov et al., 2018)) despite having significantly lower training and deployment costs.

A real-world multi-domain adaptation setting lends itself very naturally to the MDT approach. For example, domain or topic is one such dimension, whether or not the data is synthetic is another. The definition of a synthetic sample may also differ between applications. Back-translation as used in this work is an obvious way of generating such samples, but so can be pseudo-alignment (Imankulova et al., 2017; Schwenk et al., 2019). A hybrid dataset may include samples from all three origins (genuine, machine translated and pseudo-aligned) and a tag can help the model differentiate between them. Lastly, multilingual models where the source languages are not trivially different, can be boosted with a language tag[1]. It is therefore clear that although our experiments only cover a two-dimensional setting with the attributes mentioned above (data domain and source), multidimensional tagging can be extended to cover other data aspects.

4 Experimental Setup

This section describes our data sources, model architecture, and synthetic data generation and mixing strategies that we employ in our experiments. Our principal goal is to evaluate MDT fine-tuning approach as a scalable alternative to state-of-the-art domain fine-tuning for NMT.

4.1 Data

We run our experiments on three language pairs (Arabic-English, German-English and Russian-English) which span three different scripts. Our parallel data sources include a large generic corpus which is a mixture of publicly available and in-house data[2], as well as three much smaller domain-specific parallel datasets (Table 1). The monolingual data which we use to create back-translated models contains 1M proprietary text segments for each language and domain. All three domains (“Reviews”, “Messaging” and “Descriptions”) are travel-related, and in fact could be considered as sub-domains of a more general “Travel” domain. Nevertheless, they all exhibit distinct linguistic characteristics which makes it challenging to treat them as a single domain. Appendix C provides examples of sentences from different data sources.

[1]Independent experiments (not shown in this work) have shown improved results when a Portuguese model is enhanced with a tag denoting a Brazilian versus a European Portuguese author.

[2]The publicly available portion of our data was sourced from `http://opus.nlpl.eu/` Tiedemann (2012)

4.2 Synthetic data generation

We generate all synthetic data using a target-source reverse model trained purely on the generic parallel corpus. According to prior experiments we found *top10*[3] method from Edunov et al. (2018) to be the best-performing domain adaptation method, and we use it as the main approach to benchmark against. Because we do have limited in-domain parallel data, our fine-tuning parallel data is not purely synthetic, but a mix of synthetic and genuine (which we upsample to reach 1:1 composition).

top10 Following Edunov et al. (2018) we use our reverse target-source models to translate monolingual data back to English, but at the generation stage we *sample* from the next token distribution instead of using beam search to approximate MAP translation. At each sampling step we only consider top 10 most probable candidates.

MDT As described in Section 3, we extend the idea of tagged BT Caswell et al. (2019) to multi-attribute setting by prepending source-side tags which qualify various aspects of the data. Specifically, in this experiment we tag the data according to two characteristics: (1) whether it is synthetically generated or genuine, (2) which sub-domain it belongs to. Both types of tags are treated just like any other tokens, i.e. their learned embeddings are stored in the shared source-side embeddings table.

4.3 Model architecture

Prior to feeding parallel data into the sequence-to-sequence models, all text is preprocessed using the byte-pair encoding (BPE) tokenization scheme (Sennrich et al., 2016b). Our models follow the transformer-base architecture from Vaswani et al. (2017) as implemented in OpenNMT-tf[4] v1.25 (Klein et al., 2017) with early stopping based on development sets of 5000 sentences per each use case.

4.4 Evaluation

The context of this work is a real-world industrial setting which involves translating large volumes of customer-facing text. Therefore our main evaluation criteria are human-based assessments. The human evaluation was performed by professional translators on a 4-point adequacy Likert scale using 250 samples per language, per domain. Appendix A provides details of the scoring guidelines that human evaluators follow. Additionally we report case-sensitive BLEU score (Papineni et al., 2002) as implemented by sacreBLEU[5] Post (2018).

5 Results

5.1 Multi-domain adaptation

Table 2 summarizes our multi-domain adaptation results. On average MDT does not only match, but in fact outperforms the strong *top10* (Edunov et al., 2018) baseline. As mentioned in Section 4.4, given the production quality requirement of our systems we consider human scoring the gold standard for evaluating translations, not the BLEU score alone. Most human and BLEU scores do rank-wise agree, but there are some exceptions. Specifically the German-English MDT model does better than the respective *top10* models on Messaging and Descriptions domains according to the human evaluators, however it is not reflected in the BLEU scores.

[3]Our fine-tuned *top10* baseline was actually our customer-facing production system at the time for several languages.
[4]`https://github.com/OpenNMT/OpenNMT-tf`
[5]`https://github.com/mjpost/sacreBLEU`

5.2 Ablation experiment

In order to assess the role of tags, we perform an ablation experiment for German language, in which we compare the MDT performance to that of a model trained without the tags (but on the same mix of training data). It appears that the tags indeed on average improve the performance (Table 3). The models without tags perform worse on "Reviews" and "Messaging" domains according to human evaluations, and on all three domains according to the BLEU score evaluations.

	Human score	**BLEU score**
Reviews		
MDT Model	3.88	46.34
(-tags)	3.82 (-.06)	44.24 (-2.10)
Messaging		
MDT Model	3.78	49.85
(-tags)	3.48 (-.30)	49.21 (-0.64)
Descriptions		
MDT Model	3.73	50.84
(-tags)	3.80 (+.07)	49.79 (-1.05)
Average	-.10	-1.26

Table 3: The effect of tags removal on human and BLEU score in German-English MDT model.

6 Conclusions

In this work we introduce multidimensional tagging and demonstrate that it can be a scalable solution for multi-domain adaptation in a realistic resource-constrained setting. Somewhat surprisingly we find that MDT models in fact outperform on average our best alternative fine-tuning technique (*top10* from Edunov et al. (2018)), even though the alternative method trains a custom model for each sub-topic. Although the present work offers limited empirical evaluations of MDT (two dimensions: 3 sub-domains and 2 data sources; three language pairs), we think that the technique can prove useful in a broader setting. We believe it to be particularly well suited to many real-world scenarios in which practitioners develop solutions for multiple related domains, while leveraging data from different sources, both genuine and synthetic. All experimental results reported in this work follow rigorous human evaluations in addition to the standard BLEU scores assessments.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR 2015 : International Conference on Learning Representations 2015*.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the fourth workshop on statistical machine translation*, pages 182–189.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Bojar, O. and Tamchyna, A. (2011). Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336. Association for Computational Linguistics.

Britz, D., Le, Q., and Pryzant, R. (2017). Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

Chu, C. and Wang, R. (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.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., et al. (2016). Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *EMNLP 2018: 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Elaraby, M., Tawfik, A. Y., Khaled, M., Hassan, H., and Osama, A. (2018). Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6. IEEE.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Imamura, K., Fujita, A., and Sumita, E. (2018). Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 55–63.

Imankulova, A., Sato, T., and Komachi, M. (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.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Kobus, C., Crego, J., and Senellart, J. (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. INCOMA Ltd.

Levin, P., Dhanuka, N., Khalil, T., Kovalev, F., and Khalilov, M. (2017a). Toward a full-scale neural machine translation in production: the booking.com use case. In *Proceedings of MT Summit XVI*, volume 2, pages 39–49.

Levin, P., Dhanuka, N., and Khalilov, M. (2017b). Machine translation at booking. com: Journey and lessons learned. In *Proceedings of the 20th International Conference of the European Association for Machine Translation (EAMT)*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Schwenk, H., Wenzek, G., Edunov, S., Grave, E., and Joulin, A. (2019). Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (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)*, volume 1, pages 86–96.

Sennrich, R., Haddow, B., and Birch, A. (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)*, volume 1, pages 1715–1725.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5998–6008.

White, J., O'Connell, T., and O'Mara, F. (1994). The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 193–205.

Yamagishi, H., Kanouchi, S., Sato, T., and Komachi, M. (2016). Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210, Osaka, Japan. The COLING 2016 Organizing Committee.

Supplementary Material

A Human evaluations criteria

Each reported human evaluation reading is based on a random test set of 250 text samples which are evaluated by professional translators. Even though all translators were aligned and calibrated during previous evaluations, all sentences from the sample are always sent to the same individual translator to preserve consistency. We use an internally built tool (Figure 2) which allows scoring on a four-point Likert scale, a modified version of the ”Accuracy” dimension of the Fluency/Adequacy framework White et al. (1994); Callison-Burch et al. (2007); Levin et al. (2017b). We observed that fluency is almost never an issue in neural machine translation, so we do not score it explicitly. The following are the scoring guidelines for the four-point accuracy scale that are given to the translators:

4	All aspects of the review are comprehensible.
3	The fundamental information provided is accurately conveyed in the translation. Minor errors in non-essential supplementary information that are vague or obscured, but do not contend with the core of the meaning in the description, are allowed.
2	The fundamental information provided is obscured/distorted. The translation either indicates different factual information to what is present in the source, or the translation introduces incorrect information.
1	The translation does not make any sense, and/or does not even allude to the core of the source text.

B Reproducibility

Prior to feeding parallel data into the sequence-to-sequence models, all text is pre-processed using byte-pair encoding (BPE) tokenization scheme (Sennrich et al., 2016b). For all language pairs the BPE vocabulary size is set to 32k. For EN-DE language pair the vocabulary is learned jointly, while for EN-RU and EN-AR we use separate 32k vocabularies due to different alphabets in source and target. All our models follow the transformer-base architecture as described

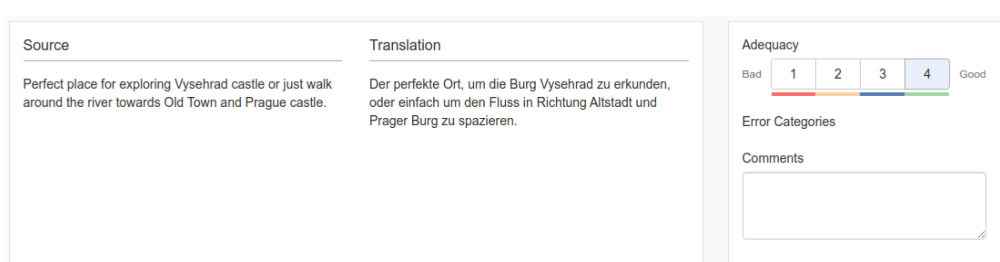

Figure 2: A screenshot of the internal human evaluation tool used by the language specialists.

in Vaswani et al. (2017) and implemented in OpenNMT-tf software (Klein et al., 2017)[6]. We trained the models using Adam Kingma and Ba (2014) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.998$ with label smoothing set to 0.1 and noam decay with an initial learning rate of 2.0. While no hyper-parameter tuning is done, early stopping is based on a dev set of 5000 sentences. Furthermore, we use an effective batch size of 25,000 tokens accumulated over different GPUs and keep training until validation loss does not decrease for two consecutive steps. We select the checkpoint with minimum sentence level validation loss - therefore completely ignoring BLEU at model selection. We report both BLEU and human evaluation results using beam width equal to four on a separate test set.

Training our base models took around 5 days using 8 NVIDIA V100 GPUs. Fine-tuning (both the single-domain baseline and the multi-domain MDT variant) took around 16 hours on a single GPU of the same model showing that there is no noticeable difference in training time. Inference time is the same for all models and only depends on sequence length.

C Text samples

The table below provides a few typical text samples from each domain for each of the three source languages. We also show English reference (human) translation as well as translation outputs from each of the three engines: base model, domain fine-tuned model (top10) and MDT (our method).

Reviews

Source	Были всего одну ночь, поэтому в полной мере оценить не смогли.
Reference	We only stayed there for one night, so we couldn't fully appreciate it.
Base model	There was only one night, so we could not fully appreciate it.
top10	We were there only for one night, so we couldn't fully appreciate it.
MDT	We were only there for one night, so we could not fully appreciate it.
Source	die Abwesenheit von Personal der Raum lies sich nicht heizen
Reference	absence of staff the room could not be heated
Base model	the absence of personnel in the room could not be heated
top10	the absence of staff the room could not be heated
MDT	the absence of staff the room could not be heated
Source	مكانه فقط
Reference	Its location only
Base model	Just his place.
top10	Its location only
MDT	Its location only

[6] https://github.com/OpenNMT/OpenNMT-tf

Messaging

Source	если можно не выше второго этажа спасибо
Reference	If possible not higher than the second floor thank you.
Base model	If you can't go above the second floor thank you
top10	if possible not higher than the second floor thank you
MDT	if possible no higher than the second floor thank you
Source	wir möchten Elli, unsere Dalmatiner Hündin mitbringen.
Reference	we would like to bring Elli, our Dalmatian dog.
Base model	We'd like to bring Elli our Dalmatian bitch.
top10	we would like to bring Elli, our Dalmatian dog.
MDT	we would like to bring our Dalmatian dog Elli.
Source	مرحبا هل الدفع بالليرة ؟ وكم التكلفة لثلاث ليالي بالليرة
Reference	Hello, is the payment in Lira? What is the cost for three nights in Lira?
Base model	Hey. Is it a lira? How much for three nights a lira?
top10	Hello! Is the payment in pounds? And how much is it for 3 nights in lira
MDT	Hello Is the payment in lira? And how much it cost for 3 nights per lira.

Descriptions

Source	Просторные апартаменты обставленные в современном стиле, но при этом по домашнему уютные.
Reference	Spacious apartments are fitted in a modern style, but are still cosy like home.
Base model	Spacious apartment with modern furnishings and homelike interiors.
top10	Spacious apartments furnished in a modern style, but at the same time homely.
MDT	Spacious apartments furnished in a modern style, but at the same time homely.
Source	Feste und Kulinarik auf höchster Ebene garantieren Abwechslung das ganze Jahr!
Reference	Festivals and culinary delights of the highest standard guarantee variety all year round!
Base model	Festive and culinary cuisine at the highest level guarantees variety all year round!
top10	Festivals and culinary delights at the highest level guarantee variety all year round!
MDT	Festivals and culinary delights at the highest level guarantee variety all year round!
Source	مكان رائع لإقامه ممتعه يقع فى قرية بورتو ساوث بيتش بالعين السخنه حيث الجو الممتع والطبيعة الخلابة وحيث يتواصل البحر بالجبل والطبيعة الخلابة
Reference	A great place for a pleasant stay located in the village of Porto South Beach in Ain Sokhna, where the atmosphere is enjoyable and picturesque nature, and where the sea meets the mountain and picturesque nature
Base model	A great place for an enjoyable stay, located in the village of Porto South Beach with the hot eye, where the atmosphere is enjoyable and nature is picturesque and where the sea communicates with the mountain and picturesque nature
top10	A great place to stay, located in the village of Porto South Beach in Ain Sokhna, where the atmosphere is pleasant and the nature is wonderful and where the sea communicates with the mountain and the wonderful nature
MDT	A great place for a pleasant stay located in the village of Porto South Beach in Ain Sokhna, where the atmosphere is pleasant and the nature is picturesque and where the sea communicates with the mountain and the picturesque nature