Improving Domain Robustness in Neural Machine Translation with Fused Topic Knowledge Embeddings

Danai Xezonaki[1,2] danai.xezonaki1@huawei-partners.com
Talaat Khalil[1*] khalil.talaat@gmail.com
David Stap[2] d.stap@uva.nl
Brandon James Denis[1] brandon.james.denis@huawei.com
[1] Huawei Technologies R&D, Amsterdam, Netherlands
[2] Language Technology Lab, University of Amsterdam

Abstract

Domain robustness is a key challenge for Neural Machine Translation (NMT). Translating text from a different distribution than the training set requires the NMT models to generalize well to unseen domains. In this work we propose a novel way to address domain robustness, by fusing external topic knowledge into the NMT architecture. We employ a pretrained denoising autoencoder and fuse topic information into the system during continued pretraining, and finetuning of the model on the downstream NMT task. Our results show that incorporating external topic knowledge, as well as additional pretraining can improve the out-of-domain performance of NMT models. The proposed methodology meets state-of-the-art on out-of-domain performance. Our analysis shows that a low overlap between the pretraining and finetuning corpora, as well as the quality of topic representations help the NMT systems become more robust under domain shift.

1 Introduction

Neural Machine Translation (NMT) has achieved impressive performance over the last few years when trained on large-scale data (Bojar et al., 2018). This success relies heavily on the availability of such data. The use of deep neural models has become the dominant approach for translation systems. However, it is not always possible to obtain neither parallel nor monolingual domain-specific data.

Most approaches for improving domain robustness in NMT assume that the target domains are known in advance, and a significant amount of data is available from the target domain. In such cases, the dominant approach for addressing domain mismatch is domain adaptation. However, when building translation systems and, as in many real-life scenarios, the target domains cannot always be known a priori. Koehn and Knowles (2017) were the first to identify domain mismatch as one of the main challenges of NMT. It is important therefore to develop translation systems that can generalize to domains unseen during training and thus be robust even under domain shift, as no target-domain data can be seen during training.

[*]Work conducted while working in Huawei.

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 209–221
September 4–8, 2023, Macau SAR, China.

Furthermore, even if NMT systems are trained on large-scale data, it is always possible that new topics or domains will emerge over time. These new domains make it difficult to maintain large translation systems, since these would require additional training on the new domains. A typical example is the outbreak of COVID-19, which intruded into everyday life and affected millions of peoples' lives. Keeping translation models up-to-date with such emerging topics is practically difficult, due to the limited availability of parallel data (Mahdieh et al., 2020).

In this work we focus on the problem of improving robustness under domain shift in NMT. A domain is defined by a corpus extracted from a specific source, and may differ from other domains in terms of topic, genre, level of formality, etc. (Koehn and Knowles, 2017). To this end, we improve domain robustness by incorporating external topic knowledge into the NMT models. We employ a denoising autoencoder that has been pretrained on Masked Language Modeling (MLM) using monolingual data and thus has not been exposed to any parallel data of the unseen test domains during training. Moreover, we train a distributional topic model using monolingual source-side data and subsequently extract a topic feature vector for each token in the vocabulary. We incorporate this external topic information during continuing the autoencoder's monolingual pretraining, and also during finetuning it on the downstream task of NMT.

To the best of our knowledge, this is the first work studying the contribution of topic modeling for domain robustness in NMT. Our key contribution is that we integrate external topic information into the NMT models, meeting state-of-the-art results for both in- and out-of-domain performance. Our analysis shows that both the quality of topic vectors and also the overlap between the pretraining and finetuning corpora are key factors towards improving domain robustness. Our results show that the proposed methodology improves domain robustness across two of the five experiments we conducted.

2 Related Work

2.1 Domain Robustness in NMT

Domain robustness has been identified as one of the main challenges of NMT (Koehn and Knowles, 2017). Müller et al. (2020) experimented with subword regularization (Kudo, 2018), defensive distillation (Hinton et al., 2015), reconstruction (Tu et al., 2017) and neural noisy channel reranking (Li and Jurafsky, 2016), and showed that reconstruction, meaning training a reconstructor component to learn to reconstruct the source sentence from decoder states, is the most effective technique for improving out-of-domain robustness in NMT.

In addition, Wang and Sennrich (2020) correlated domain robustness with hallucinations and proposed Minimum Risk Training, a sentence-level training objective, in order to reduce hallucinations and thereby improve indirectly domain robustness. Müller and Sennrich (2021) further examined the role of Minimum Bayes Risk Decoding and showed that it can indeed increase the robustness against domain shift. Moreover, Berard et al. (2020) found that initializing the NMT encoder using pretrained embeddings from language models helped out-of-domain robustness, while Germann (2020) proposed improving robustness by adding noise to the output layers of the NMT systems.

2.2 Pretraining in NMT

Unsupervised pretraining has been widely used over the last years, in order to deal with scarcity of large parallel in-domain data. It has been shown that pretraining acts as a regularizer in deep neural networks, and thus allows better generalization (Erhan et al., 2010). During pretraining, large models are typically trained with a denoising objective using monolingual data, as Masked Language Modeling (MLM), and are subsequently finetuned on downstream NLP tasks.

Recent studies have shown that pretrained NLP models can further improve out-of-domain

robustness in NMT (Hendrycks et al., 2020; Tu et al., 2020). However, Liu et al. (2021) claimed that MLM training teaches the decoder to copy tokens from the input to the output of the system, and addressed this limitation by proposing a copying penalty, which mitigates the copying behavior of NMT systems. Through their experiments, they showed that the proposed method was able to improve even out-of-domain robustness.

2.3 Topic Modeling

Topic modeling has also been employed in the context of NMT, in order to provide prior semantic knowledge to the models. Topic models are statistical tools which identify hidden patterns and semantic structure in text corpora (Blei et al., 2010). Despite the fact that it has been shown that topic models significantly improve translation performance when incorporated into NMT architectures (Zhang et al., 2016; Chen et al., 2016; Wang et al., 2021), it has been yet unexplored how they can contribute to domain robustness in NMT. In this work, we go a step further and show that external topic information can also improve the lexical selection of the NMT systems under domain shift and thus help them become more domain robust.

In contrast to statistical topic models, various works have proposed distributional topic algorithms that mix Latent Dirichlet Allocation (Blei et al., 2001) with word embeddings (Mikolov et al., 2013a,b). Dieng et al. (2020) proposed the Embedded Topic Model (ETM), which is used in this work. ETM is a generative probabilistic model which assumes that each word is modeled by a categorical distribution and each document is a mixture of topics. The words are represented by an embedding, and the topics are points in the same embedding space. The distribution of topics over words is then defined by the dot product between each word and each topic embedding.

3 MBARTOPIC

In this section we introduce MBARTOPIC, our proposed system in order to improve domain robustness in NMT. We employ a sequence-to-sequence system and initialize its weights using a pretrained model. We need to ensure that the pretrained model has not been exposed to any parallel data of the test domains. To this end, we use a multilingual denoising sequence autoencoder for initialization, and, in particular mBART (Liu et al., 2020), which has been trained on monolingual data only, and on a different task than NMT.

Assuming a corpus D consisting of sub-corpora D_j , where each D_j is a set of monolingual text samples $D_j = (X_1, X_2, ..., X_n)$, as well as a noising function g that corrupts text, the mBART model is trained to reconstruct X_i from $g(X_i)$. Therefore, during pretraining it learns to maximize the log-likelihood of predicting the input X, given a noisy variant of it, as follows:

$$L = \sum_{D_j \in D} \sum_{X_i \in D_j} \log p(X_i | g(X_i); \theta) \tag{1}$$

We employ this pretrained system and finetune it on the downstream NMT task. We feed the parallel domain-specific data to the encoder and the decoder, while also adding the special ID token of the source and target languages, respectively. During finetuning on NMT the model learns to minimize the cross-entropy loss function:

$$L_{CE} = -\sum_{t=1}^{n} \log p_\theta(y_t | y_{<t}, x), \tag{2}$$

where x is the input sequence, y is the generated output and y_t is the t-th generated token.

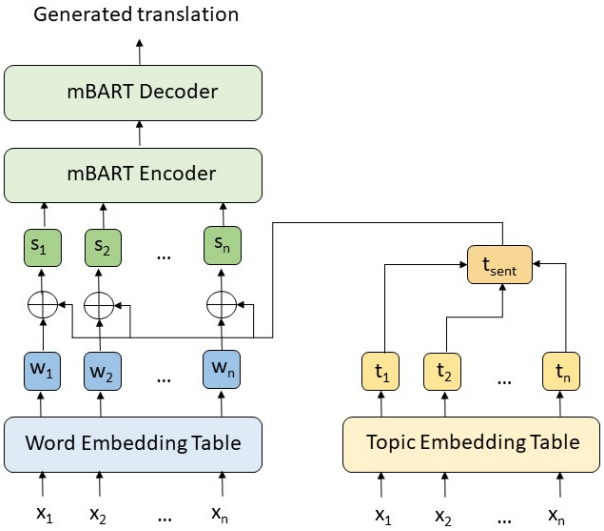

Figure 1: The illustration of the proposed MBARTOPIC architecture. The $\oplus$ is a sum operation. The t_{sent} is the sentence-level topic information obtained from the source tokens and the process of computation is given by Equation 3.

3.1 Integrating External Topic Information

A domain may differ from other domains in terms of topic, genre, level of formality, etc. (Koehn and Knowles, 2017). Based on this definition, we employ external topic information and incorporate it to the NMT models. The proposed system is shown in Figure 1.

We obtain the external topic knowledge by training a distributional topic model and extracting its topic embedding tables. Specifically, we employ the Embedded Topic Model (ETM, Dieng et al. (2020)) and train it on large monolingual corpora of the source language. Subsequently, we freeze the topic model and we use its trained topic embeddings. We extract one feature vector t_i for every term in the input sequence, which serves as an external context vector. We follow Wang et al. (2021) by employing their 'ENC_{pre}' topic integration method and propose an experimental setup where the extracted topic vectors can be utilized in the low-resource domain robustness scenario for NMT. As shown in Figure 1, we incorporate the topic vectors to the model architecture, by adding them to the embedding vector of each input token. In particular, we take the average of all the topic vectors of the source tokens and pass it through a projection layer. Through this projection we extract a sentence-level topic representation, t_{sent}, as follows:

$$t_{sent} = f\left(\frac{1}{n}\sum_{i=1}^{n} t_i\right), \tag{3}$$

where f is a learnable mapping.

The sentence-level topic information is then added to the word embeddings of the input tokens. Therefore, the final input representation for the i-th input token is given by:

$$s_i = w_i + t_{sent}, \tag{4}$$

where w_i is the embedding of the i-th input token. This combined input is finally fed to the encoder of mBART.

domains	corpora	size
IT	GNOME, KDE, PHP, Ubuntu, OpenOffice	$222,927$
Law	JRC-Acquis	$467,309$
Medical	EMEA	$248,099$
Koran	Tanzil	$17,982$
Subtitles	OpenSubtitles2018	$500,000$

Table 1: Dataset overview. Size indicates number of sentence pairs after filtering.

We experiment with both continuing the pretraining of mBART, as well as finetuning it. During the continued pretraining, we train mBART on denoising source-language monolingual data. During finetuning, we finetune mBART on domain-specific parallel data. In both cases, we incorporate the external topic information to the source-side data as shown in Figure 1. The final model is then evaluated on translating out-of-domain data.

4 Experimental Setup

We compare five different systems:

1. **BASELINE**: As a weak baseline, we train a Transformer Base model (Vaswani et al., 2017).
2. **RANDOMINIT**: We train the mBART-large architecture from scratch, on domain-specific data. This experiment differs from MBART-FT as here we do not employ the pretrained weights of mBART, but instead initialize the network weights randomly. This model serves towards evaluating the contribution of pretraining.
3. **MBART-FT**: We employ the pretrained mBART-large system[1] and we do standard finetuning on our parallel data. Finetuning is performed on one domain at a time, and the models are evaluated on both seen and unseen domains.
4. **MBART-PT-FT**:We continue the pretraining and finetune mBART-large, but without adding any topic information at all. This experiment serves towards comparing against MBARTOPIC and thus discriminating the contribution of topics from the contribution of pretraining the model for more gradient updates.
5. **MBARTOPIC**: We augment the MBART-FT model with external topic knowledge, which is fed during additional pretraining of mBART and during finetuning on domain-specific data.

4.1 Datasets

We report experiments in the German $\rightarrow$ English (DE $\rightarrow$ EN) language direction. To verify the effectiveness of the proposed methods, we use corpora from five distant domains: IT, Medical, Law, Koran and Subtitles. For all experiments we make use of the same data as Müller et al. (2020); Wang and Sennrich (2020); Liu et al. (2021), as made available from the OPUS collection (Tiedemann, 2012). Each domain contains 2000 sentence pairs for evaluation and 2000 for testing. Additional details about the specific datasets of each domain and their sizes are shown in Table 1. For each experiment, we use one domain for pretraining/finetuning the models, and all five domains for testing both in- and out-of-domain performance.

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/mbart/README.md

As monolingual source-side data for training the topic model and also for continuing the pretraining of mBART, we employ generic monolingual data from the news domain, and specifically the German News Dataset (Mi, 2020). This dataset is a collection of around $175k$ newspaper articles in German, where the articles are extracted from 15 news websites.

4.2 Preprocessing

For the BASELINE system, we use a joint BPE vocabulary (Sennrich et al., 2016) which is learnt with $32k$ merge operations over the entire corpus, taking both source and target samples into account. We preprocess the data by applying tokenization, normalizing punctuation, cleaning and removing non-printing characters using Moses (Koehn et al., 2007). For continuing the pretraining and finetuning, we do the same bpe processing as in the rest experiments, but we use the mBART pretrained sentencepiece tokenizer. We also use the same tokenizer to tokenize the monolingual data that we used to train the topic model.

4.3 Implementation Details

We implemented all experiments using FAIRSEQ (Ott et al., 2019). Our models use the Transformer architecture (Vaswani et al., 2017). Models are trained for a maximum of $100k$ steps with 1024 maximum tokens per GPU. All models are trained using 8 Nvidia Tesla-V100 GPUs. The continued pretraining, finetuning and topic model training required approximately 280, 470 and 210 GPU hours respectively. The Transformer Base systems consist of around $60M$ parameters and the mBART-based systems consist of approximately $610M$ parameters. We decode using beam search and a beam size of 5 and a length penalty of 1.4. Similar to the related works, we report case-sensitive BLEU (Papineni et al., 2002) scores on detokenized text using `sacrebleu` (Post, 2018).

We optimize all models with Adam (Kingma and Ba, 2015). We use early stopping to choose the model with the lowest loss on the validation set. For the baseline experiments, we use 5^{-4} maximum learning rate, 4000 warm-up steps and 0.2 dropout. For continuing the pretraining we mask 35% of the input words and train with 0.3 dropout. For finetuning, we train with 0.2 dropout. We also use 0.2 label smoothing, 2500 warm-up steps, polynomial decay and 3^{-5} maximum learning rate for both finetuning and continued pretraining.

For the topic model, we use the Embedded Topic Model (Dieng et al., 2020). We train the model for 500 epochs and set the number of topic clusters to 50 as in Wang et al. (2021). The embedding dimension of the trained topic vectors is set to 300.

5 Results

For each experiment, we train all systems on one domain at a time and evaluate them on the same (in-domain) and also on the rest four domains (out-of-domain). In Table 2 we compare the performance of the proposed models. In each sub-table of results, we report the train domains vertically and the test domains horizontally.

We observe that training the Transformer Base model (BASELINE) yields better results across almost all experiments, compared to training the mBART architecture from scratch (RANDOMINIT). This finding is expected given the large difference in the number of parameters between the two models and the relatively small amount of data we used to train the systems from scratch.

Moreover, initializing the network with pre-fitted weights (MBART-FT) is shown to achieve a significant gain in performance, compared to the results of RANDOMINIT and BASELINE, across all experiments, for both in- and out-of-domain translation. This observation

stands in agreement with related literature indicating improvements in downstream NLP tasks when initializing the models with pretrained weights, due to transferring general knowledge to them (Liu et al., 2021), which contributes positively towards out-of-domain robustness.

		Test domains						
System		IT	Law	Koran	Medical	Subtitles	**OOD p.d.**	**Avg. OOD**
BASELINE	IT	42.5	9.7	2.3	16.9	8.4	9.3	7.5
	Law	15.8	60.0	2.0	24.2	5.5	11.9	
	Koran	0.2	0.2	14.6	0.1	1.0	0.4	
	Medical	12.4	15.7	1.5	57.1	4.6	8.6	
	Subtitles	8.1	5.1	5.8	9.7	21.3	7.2	
RANDOMINIT	IT	34.8	4.6	1.5	5.6	6.3	4.5	3.5
	Law	4.4	45.2	1.1	8.1	2.3	4.0	
	Koran	0.3	0.3	14.3	0.3	0.7	0.4	
	Medical	4.1	9.3	1.2	47.7	2.2	4.2	
	Subtitles	5.6	3.0	4.6	4.6	23.9	4.5	
MBART-FT	IT	58.2	21.4	4.9	28.1	14.5	**17.2**	14.3
	Law	28.3	76.3	2.6	30.9	7.5	17.3	
	Koran	0.8	1.2	19.2	1.1	3.1	**1.6**	
	Medical	26.0	25.6	2.0	66.0	7.3	15.2	
	Subtitles	29.0	18.6	6.4	26.8	26.4	**20.2**	
MBART-PT-FT	IT	59.3	20.1	4.5	26.8	13.4	16.2	14.6
	Law	29.1	76.3	2.8	32.0	7.5	**17.9**	
	Koran	0.4	1.0	18.9	0.9	2.7	1.3	
	Medical	30.7	31.6	2.0	58.1	7.4	17.9	
	Subtitles	27.1	18.1	6.6	27.2	25.8	19.8	
MBARTOPIC	IT	59.6	20.3	4.6	27.4	13.3	16.4	14.7
	Law	29.1	76.3	2.7	31.0	7.7	17.6	
	Koran	0.5	0.9	18.7	0.6	2.6	1.2	
	Medical	30.7	31.4	2.3	58.0	8.4	**18.2**	
	Subtitles	27.2	18.3	6.7	28.0	26.1	20.1	

Table 2: Case-sensitive BLEU results of the 5 models for DE→EN. 'OOD p.d.' stands for the averaged out-of-domain score per domain-experiment. 'Avg. OOD' stands for the total average out-of-domain score, per system. We highlight the highest out-of-domain score per domain experiment in bold.

Comparing the MBART-PT-FT system to not performing any additional pretraining, as in MBART-FT, it becomes evident that the extra pretraining updates improve the domain robustness of the Law and Medical experiments. These models have a significant increase in their out-of-domain performance, while the Medical system seems to become more general after continuing the pretraining, given the decrease to its in-domain score. On the other hand, additional pretraining seems not to help the performance of the IT, Koran and Subtitles experiments. We argue in Section 6.1 that this finding is related to the overlap between the monolingual corpus used for pretraining and the finetuning domains.

Furthermore, we compare MBARTOPIC to MBART-PT-FT and observe that incorporating the topic knowledge to the system improves the out-of-domain translation for the IT, Medical and Subtitles experiments. On the other hand, the topic knowledge fusion has a slight decrease in the robustness of the Law and Koran experiments.

We would like to point out that for continuing the pretraining and for training the topic model, we used monolingual data from the news domain, which are relatively small and does not generalize well for all possible topics. These data are also biased towards specific domains, as per our overlap study in Section 6.2. We hypothesize that using large scale web crawled data for learning topic embeddings and continuing pretraining, e.g the mBART pre-training data, might help get rid of the side effect of knowledge overriding as suggested by the analysis, resulting in higher quality topic representations. We leave the investigation of this hypothesis to future work, since we were limited by computational constraints.

Finally, we provide some example translations of our systems in the out-of-domain setting [2].

System	Medical ID	Avg. Medical OOD
SMT (Müller et al., 2020)	58.4	11.8
NMT (Müller et al., 2020)	61.5	11.7
NMT+RC+SR+NC (Müller et al., 2020)	60.8	13.1
MLE w/ LS + MRT (Wang and Sennrich, 2020)	58.8	12.0
PRETRAINED (Liu et al., 2021)	63.1	17.6
PRETRAINED + CP (Liu et al., 2021)	**63.2**	**18.3**
MBART-FT	**66.0**	15.2
MBART-PT-FT	58.1	17.9
MBARTOPIC	58.0	**18.2**

Table 3: BLEU scores for the Medical experiment. We compare in-domain (ID) and out-of-domain (OOD) performance.

5.1 Comparison to Related Work

In Table 3 we compare our systems to the Related Work. These works employ the same corpora as we did, train their models on the Medical domain and evaluate them on all 5 domains. To this end, we compare our results of the Medical domain. We can see that our proposed systems perform comparably to the best systems in terms of out-of-domain performance, with Liu et al. (2021) achieving 18.3 and our MBARTOPIC model achieving 18.2 BLEU score. We also note that our proposed methodology might be orthogonal to the PRETRAINING + CP model of Liu et al. (2021); therefore combining them may lead to additional increases in quality.

Our MBART-FT model additionally achieves a high in-domain score. It should be noted that the PRETRAINED experiment (Liu et al., 2021) is the same experiment as our MBART-FT model. We attempted to replicate the results of PRETRAINED but unfortunately we were unable to do so. We were not able to determine the reason for the discrepancy. Overall, our most domain robust system for the Medical domain is MBARTOPIC, which achieves the biggest improvement in terms of out-of-domain performance.

6 Contributing Factors to Domain Robustness

To understand what contributes to the out-of-domain performance gains, we conduct an analysis of the results presented in Section 5.

6.1 N-gram Overlap between Pretraining and Finetuning Corpora

Recall that Table 2 shows that additional pretraining helps the Law and Medical domains achieve the most significant improvement in their out-of-domain scores. We hypothesize that

[2]https://gist.github.com/danaiksez/7cfe3463ebf43b188e37689c104075d2

Domains	Uni-grams (%)	Bi-grams (%)	Tri-grams (%)	Four-grams (%)
IT	96.58	4.74	0.98	0.18
Law	**84.65**	2.88	1.38	0.20
Koran	99.52	5.92	1.09	0.08
Medical	**91.18**	3.89	0.92	0.18
Subtitles	97.63	2.83	1.61	0.24

Table 4: Overlap of n-grams between the n-gram vocabularies of the German News monolingual dataset and each of the 5 domain corpora. The values are calculated using Equation 5.

this finding is related to the topic and style of the monolingual data used for additional pretraining. We investigate this by analyzing how much related is the pretraining to the finetuning datasets. To this end, we compute the overlap of n-grams between the German News Dataset and each of the five domain corpora. We particularly consider the uni-, bi-, tri- and four-gram vocabularies of the corpora we employ and calculate the percentage of the German News n-grams that co-exist in the finetuning domain vocabularies. Those n-gram overlaps are shown in Table 4. The values are calculated as the percentage of the following:

$$overlap_{ngram} = \frac{\#shared\_ngrams}{\#pretraining\_ngrams}, \tag{5}$$

where $\#shared\_ngrams$ is the number of shared entries between the pretraining and finetuning n-gram vocabularies, and $\#pretraining\_ngrams$ is the n-gram vocabulary size of the pretraining (German News) dataset.

We observe that the Law and Medical experiments, which improved their out-of-domain performance through additional pretraining, have the lowest overlap with the pretraining corpus. This finding suggests that during pretraining, the systems acquire extra general knowledge through the denoising of monolingual corpora. Therefore, during finetuning, when there is smaller n-gram overlap with the pretraining dataset, it is likely that the previously acquired general semantic knowledge helps the systems generalize well to unseen domains.

On the other hand, in the case of larger overlap, the models seem to 'erase' this general knowledge they aquired during pretraining, and instead overwrite it with the domain-specific context they are exposed to, during finetuning. To this end, these systems are likely to overfit on the finetuning domain and forget about the more general information. This in turn affects negatively their translation robustness under domain shift.

6.2 Topic Embeddings Analysis

We analyze the topic embedding vectors and their contribution when incorporated to the system architecture. We do that by measuring the distances between the intra-domain trained topic vectors.

We assume that the most important words per domain should lie closer to each other in the embedding space. In this case, fusing the topic representations to the network architecture should contribute towards discriminating the domains easier. We choose the top-n most 'categorical' words per domain. These are selected as the ones with the highest TF-IDF score. We then measure for each domain, the cosine distance between each possible pair of them. Since not all words are equally significant for a domain corpus, we want the final score to reflect the importance of the words. Therefore, we weigh the cosine distance of each word pair by the IDF scores of both words.

Given the top-n words of a specific domain, with n empirically selected to 10 in our case, and w_i being the i-th word of the top-n words, each entry of Table 5 is computed as follows:

$$\text{WCD} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \cos_D(w_i, w_j) \text{IDF}(w_i) \text{IDF}(w_j) \tag{6}$$

Table 5 shows the averaged Weighted Cosine Distances (WCD) per domain. We notice that the WDC between the most important words of Law and Koran are the highest among the domains. As shown in Table 2, the Law and Koran experiments experience a decrease in performance when topics are added. On the other hand, the IT, Medical and Subtitles corpora, which have a smaller WCD among their most categorical words, have an increase in out-of-domain performance when fusing topic representations. These findings therefore seem to correlate with the behaviour of the experiments when fusing topic information, and highlight the need for a more concise topic embedding space.

	IT	Law	Koran	Medical	Subtitles
WCD	1.071	**1.202**	**1.347**	1.069	0.941

Table 5: Averaged Weighted Cosine Distances (WCD) between top-10 most categorical words per domain, weighted by their IDF score, as shown in Equation 6.

7 Conclusion

In this work we propose MBARTOPIC, a novel model for improving domain robustness in NMT with integrated external topic knowledge. This is the first work studying the contribution of topic information towards improving domain robustness in NMT. We use a sequence-to-sequence model, and specifically a pretrained multilingual denoising autoencoder. We train a distributional topic model on source-side monolingual data and integrate this topic knowledge to the encoder of the NMT system. We do that by extracting sentence-level topic features and subsequently combining them with the word embeddings of the each input token. In our approach we continue the pretraining of the denoising model using source-side monolingual corpora, and then finetune it on the downstream NMT task, using domain-specific parallel data. We incorporate the external topic features into both the additional pretraining and also during finetuning.

Our results show that the proposed method can improve the domain robustness of our experiments and meets state-of-the-art results in the out-of-domain performance. Our analysis suggests that additional self-supervised pretraining with a low overlap between the pretraining and finetuning corpora can be an important factor to the domain robustness of NMT systems. Finally, we show that smaller distances among the topic vectors of domain-specific words result in an increase in the out-of-domain performance.

In the future, we plan to investigate the contribution of topic knowledge when it is fused into both the encoder and decoder of the NMT system. We also plan to analyze the system performance on a leave-one-out scenario, when finetuned on multiple domains and evaluated on an unseen one.

Acknowledgements

We would like to thank Fokko Beekhof, Jun Luo and Nikolas Zygouras for their valuable suggestions and comments.

References

Berard, A., Calapodescu, I., Nikoulina, V., and Philip, J. (2020). Naver labs Europe's participation in the robustness, chat, and biomedical tasks at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 462–472, Online. Association for Computational Linguistics.

Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6):55–65.

Blei, D., Ng, A., and Jordan, M. (2001). Latent dirichlet allocation. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

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.

Chen, W., Matusov, E., Khadivi, S., and Peter, J.-T. (2016). Guided alignment training for topic-aware neural machine translation. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 121–134, Austin, TX, USA. The Association for Machine Translation in the Americas.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(19):625–660.

Germann, U. (2020). The University of Edinburgh's submission to the German-to-English and English-to-German tracks in the WMT 2020 news translation and zero-shot translation robustness tasks. In *Proceedings of the Fifth Conference on Machine Translation*, pages 197–201, Online. Association for Computational Linguistics.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. (2020). Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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.

Koehn, P. and Knowles, R. (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.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Li, J. and Jurafsky, D. (2016). Mutual information and diverse decoding improve neural machine translation.

Liu, X., Wang, L., Wong, D. F., Ding, L., Chao, L. S., Shi, S., and Tu, Z. (2021). On the copying behaviors of pre-training for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4265–4275, Online. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mahdieh, M., Chen, M. X., Cao, Y., and Firat, O. (2020). Rapid domain adaptation for machine translation with monolingual data. *ArXiv*, abs/2010.12652.

Mi, S. (2020). German news dataset.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Müller, M., Rios, A., and Sennrich, R. (2020). Domain robustness in neural machine translation. In *14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, pages 151–164. Association for Machine Translation in the Americas.

Müller, M. and Sennrich, R. (2021). Understanding the properties of minimum Bayes risk decoding in neural 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 1: Long Papers)*, pages 259–272, Online. Association for Computational Linguistics.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (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.

Tiedemann, J. (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, Istanbul, Turkey. European Language Resources Association (ELRA).

Tu, L., Lalwani, G., Gella, S., and He, H. (2020). An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Tu, Z., Liu, Y., Shang, L., Liu, X., and Li, H. (2017). Neural machine translation with reconstruction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, C. and Sennrich, R. (2020). On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Wang, W., Peng, W., Zhang, M., and Liu, Q. (2021). Neural machine translation with heterogeneous topic knowledge embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3197–3202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhang, J., Li, L., Way, A., and Liu, Q. (2016). Topic-informed neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1807–1817, Osaka, Japan. The COLING 2016 Organizing Committee.