Domain Robustness in Neural Machine Translation

Domain Robustness in Neural Machine Translation

Mathias Müller mmueller@cl.uzh.ch
Annette Rios rios@cl.uzh.ch
Rico Sennrich sennrich@cl.uzh.ch
Department of Computational Linguistics, University of Zurich, Switzerland

Abstract

Translating text that diverges from the training domain is a key challenge for machine translation. Domain robustness—the generalization of models to unseen test domains—is low for both statistical (SMT) and neural machine translation (NMT). In this paper, we study the performance of SMT and NMT models on out-of-domain test sets. We find that in unknown domains, SMT and NMT suffer from very different problems: SMT systems are mostly adequate but not fluent, while NMT systems are mostly fluent, but not adequate. For NMT, we identify such hallucinations (translations that are fluent but unrelated to the source) as a key reason for low domain robustness. To mitigate this problem, we empirically compare methods that are reported to improve adequacy or in-domain robustness in terms of their effectiveness at improving domain robustness. In experiments on German→English OPUS data, and German→Romansh (a low-resource setting) we find that several methods improve domain robustness. While those methods do lead to higher BLEU scores overall, they only slightly increase the adequacy of translations compared to SMT.

1 Introduction

Even though neural models have improved the state-of-the-art in machine translation considerably in recent years, they still underperform in specific conditions. One such condition is out-of-domain translation. Koehn and Knowles (2017) found that neural machine translation (NMT) systems perform poorly in such settings and that their poor performance cannot be explained solely by the fact that out-of-domain translation is difficult: non-neural, statistical machine translation (SMT) systems were superior at this task. For this reason, Koehn and Knowles (2017) identified translation of out-of-domain text as a key challenge for NMT.

Catastrophic failure to translate out-of-domain text can be viewed as overfitting to the training domain, i.e. systems learn idiosyncrasies of the domain rather than more general features. Our goal is to learn models that generalize well to unseen data distributions, including data from other domains. We will refer to this property of showing good generalization to unseen domains as **domain robustness**.

We consider domain robustness a desirable property of NLP systems, along with other types of robustness, such as robustness against adversarial examples (Goodfellow et al. 2015) or typos in the input (Belinkov and Bisk, 2018). While domain adaptation with small amounts of parallel or monolingual in-domain data has proven very effective for NMT (e.g. Luong and Manning, 2015; Sennrich et al., 2016a; Kobus et al., 2017; Li et al., 2019), the target domain(s) may be unknown when a system is built, and there are language pairs for which training data is only available for limited domains. Hence, domain robustness of systems without any domain adaptation is not only of theoretical interest, but also relevant in practice.

SRC *Aber geh subtil dabei vor.*

REF *But be subtle about it.*

HYP *Pharmacokinetic parameters are not significantly affected in patients with renal impairment (see section 5.2).*

Figure 1: Example illustrating how a German→English NMT system trained on medical text *hallucinates* the translation of an out-of-domain input sentence.

Model architectures and training techniques have evolved since Koehn and Knowles (2017)'s study, and it is unclear to what extent this problem still persists. We therefore revisit the hypothesis that NMT systems exhibit low domain robustness. In preliminary experiments, we demonstrate that current models still fail at out-of-domain translation: BLEU scores drop drastically for test domains other than the training domain. However, the overall out-of-domain translation quality of NMT systems is now on par with SMT systems.

An analysis of our baseline systems reveals that **hallucinated** content occurs frequently in out-of-domain translations (see Figure 1 for an example). Several authors present anecdotal evidence for NMT systems occasionally falling into a *hallucination* mode where translations are grammatically correct but unrelated to the source sentence (Arthur et al., 2016; Koehn and Knowles, 2017; Nguyen and Chiang, 2018). Our manual evaluation shows that hallucination is more pronounced in out-of-domain translation. We therefore expect methods that alleviate the hallucination problem to indirectly improve domain robustness.

As a means to reduce hallucination, we experiment with several techniques and assess their effectiveness in improving domain robustness: reconstruction (Tu et al., 2017; Niu et al., 2019), subword regularization (Kudo, 2018), neural noisy channel models (Li and Jurafsky, 2016; Yee et al., 2019), and defensive distillation (Papernot et al., 2016), as well as combinations of these techniques. The main contributions of this paper are:

- we perform an analysis of SMT and NMT systems that confirms that while in-domain BLEU increased, domain robustness remains a major problem even with state-of-the-art Transformer architectures. Our comparison of SMT and NMT shows differences in how performance degrades in unseen domains: SMT mostly suffers in terms of fluency, while NMT tends to produce more fluent, but less adequate translations (hallucinations).
- we test several techniques related to adequacy, robustness, or out-of-domain translation in regard to their effectiveness in improving domain robustness in NMT. We find that several techniques are moderately successful, most notably reconstruction, which reduces the average percentage of hallucinations in out-of-domain test sets from 35% to 29%.
- we show that despite moderate improvements, domain robustness remains a challenge in NMT, and provide code and data sets to serve as baselines for future work.

2 Data Sets

We report experiments on two different translation directions: German→English (DE→EN) and German→Romansh (DE→RM).

2.1 German→English

For all DE→EN experiments, we use the same corpora as Koehn and Knowles (2017), available from OPUS (Lison and Tiedemann, 2016).

DE–EN			DE–RM		
domains	corpora	size	domains	corpora	size
medical	EMEA	1.1m	law	Allegra,	
IT	GNOME, KDE, PHP, Ubuntu, OpenOffice	380k		Press Releases	100k
koran	Tanzil	540k	blogs	Convivenza	20k
law	JRC-Acquis	720k			
subtitles	OpenSubtitles2018	22.5m			

Table 1: Data sets common to all of our experiments. Size indicates number of sentence pairs.

We use corpora from OPUS to define five domains: *medical*, *IT*, *koran*, *law* and *subtitles*. See Table 1 for an overview of sizes per domain. The domains are quite distant, and we therefore expect that systems trained on a single domain will have low domain robustness if tested on other domains.

For each domain, we select 2000 consecutive sentence pairs each for development and testing. Our test sets are different from Koehn and Knowles (2017), so results are not directly comparable. In all experiments, the *medical* domain serves as the training domain, while the remaining four domains are used for testing.

2.2 German→Romansh

To complement our DE→EN experiments, we also train systems for DE→RM. Romansh is a Romance language that, with an estimated $40\,000$ native speakers, is low-resource, but has some parallel resources thanks to its status as an official Swiss language. Domain robustness is of particular relevance in low-resource settings since training data is typically only available for few domains. Our training data consists of $100\,000$ sentence pairs, specifically the Allegra corpus by Scherrer and Cartoni (2012) which contains mostly law text, and an in-house collection of government press releases. As test domain (unseen during training), we use blog posts from Convivenza[1]. From both data sets we randomly select 2000 consecutive sentence pairs as test sets.

3 State-of-the-art Models Exhibit Low Domain Robustness

In this section, we establish that current NMT systems exhibit low domain robustness by analyzing our baseline systems automatically and manually.

3.1 Experimental Setup for Baseline Models

We use Moses scripts for punctuation normalization and tokenization. We apply truecasing trained on in-domain training data. Similarly, we apply BPE (Sennrich et al., 2016b) with 32k (DE→EN) and 16k (DE→RM) merge operations learned from in-domain data. We train two baselines:

NMT Baseline A standard Transformer base model trained with Sockeye (Vaswani et al., 2017; Hieber et al., 2018).

SMT Baseline A standard, phrase-based statistical model trained with Moses (Koehn et al., 2007), using mtrain (Läubli et al., 2018) as frontend with standard settings.

[1] https://www.suedostschweiz.ch/blogs/convivenza

	SMT	NMT
in-domain		
medical	58.4	**61.5**
out-of-domain		
IT	**21.4**	17.1
koran	**1.4**	1.1
law	19.8	**25.3**
subtitles	**4.7**	3.4
average (out-of-domain)	**11.8**	11.7

(a)

	SMT	NMT
in-domain		
law	45.2	**52.5**
out-of-domain		
blogs	15.5	**18.9**

(b)

Table 2: BLEU scores of (a) baseline DE→EN systems trained on *medical* data, (b) baseline DE→RM systems trained on *law* data.

We always test on several domains, including the training domain. We use a beam size of 10 to translate test data. We report case-sensitive BLEU (Papineni et al., 2002) scores on detokenized text, computed with SacreBLEU (Post, 2018)[2].

3.2 Analysis of Baseline Systems

Tables 2a and 2b show automatic evaluation results for all our baseline models. Neural models achieve good performance on the respective in-domain test sets (61.5 BLEU on *medical* for DE→EN; 52.5 BLEU on *law* for DE→RM), but on out-of-domain text, translation quality is clearly diminished, with an average BLEU of roughly 12 (DE→EN) and 19 (DE→RM). The following analysis will focus on our DE→EN baseline systems.

Compared to results reported by Koehn and Knowles (2017), NMT has improved markedly since their study was conducted, and is now on par with SMT in out-of-domain settings (11.8 BLEU versus 11.7). However, on the in-domain test set, our NMT baseline outperforms the SMT baseline by 3 BLEU. This result suggests that higher in-domain performance does not guarantee better out-of-domain translations.

Unknown words constitute one possible reason for failing to translate out-of-domain texts. As shown in Table 3a, the percentage of words that are not seen during training is much higher in all out-of-domain test sets. However, unknown words cannot be the only reason for low translation quality: The test sets with the lowest BLEU scores (*koran* and *subtitles*) actually have an out-of-vocabulary (OOV) rate similar to the *IT* test set, where BLEU scores are much higher for both baseline models.

Additionally, our SMT baseline shows better generalization to some domains unseen at training time, while the average BLEU is comparable to the NMT baseline. In the *IT* domain, the result is most extreme: the SMT system beats the neural system by 4.3 BLEU. This demonstrates that the low domain robustness of NMT is not (only) a data problem, but also due to the model's inductive biases.

As a further control, we train additional baseline systems trained on *all* domains.[3] We use it to test whether the data we have held out for out-of-domain testing is inherently more difficult to translate than the in-domain test set. The results in Table 3b show that this is not the case.

[2]SacreBLEU version signature: `BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.1`.

[3]The *subtitles* domain (23m sentences) was subsampled to 1m sentence pairs so as not to overwhelm the remaining domains (3m sentences in total). We also removed any overlap between the training and test sets.

	OOV rate
in-domain	
medical	2.42%
out-of-domain	
IT	20.09%
koran	18.63%
law	9.39%
subtitles	18.16%

(a)

	SMT	NMT
in-domain		
medical	53.2	**61.4**
out-of-domain		
IT	31.9	**44.7**
koran	15.2	**15.9**
law	58.3	**62.3**
subtitles	14.9	**18.5**
average (out-of-domain)	30.1	**35.4**

(b)

Table 3: For the language pair DE→EN: (a) Out-of-vocabulary (OOV) rates of in-domain and out-of-domain test sets. (b) BLEU scores of baselines trained on a concatenation of *all* domains.

For the NMT baseline, BLEU ranges between 15.9 and 62.3, with an average out-of-domain BLEU of 35.4.

3.2.1 Hallucination

NMT models can be understood as language models of the target language, conditioned on a representation of a source text. This means that NMT models have no explicit mechanism—as SMT models do—that enforces coverage of the source sentence, and if the representation of an out-of-domain source sentence is outside the training distribution, it can be seemingly ignored. This gives rise to a tendency to *hallucinate* translations, i.e. to produce translations that are fluent, but unrelated to the content of the source sentence (Lee et al., 2018).

We hypothesize that hallucination is more common in out-of-domain settings. A small manual evaluation performed by the main author confirms that this is indeed the case. We evaluate the fluency and adequacy of our baselines (we refer to them as **NMT** and **SMT**). In a blind setup, we annotate a random sample of 100 sentence pairs per domain. As controls, we mix in pairs of `(source, actual reference)`, treating the reference translation as an additional system.

Evaluation of adequacy The annotator is presented with a sentence pair and asked to judge whether the translation is adequate, partially adequate or inadequate.

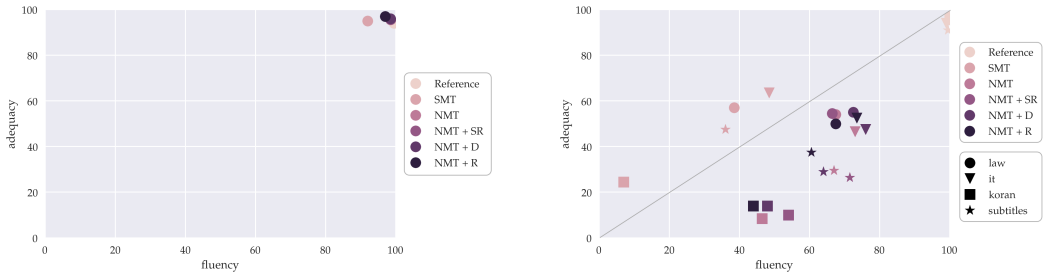
Evaluation of fluency We use the same data as for the evaluation of adequacy, however, the annotator is shown only the translation, without the corresponding source sentence. The annotator is asked whether the given sentence is fluent, partially fluent or not fluent.

Figure 2 shows the results of the manual evaluation in terms of adequacy and fluency. For visualization, individual fluency values are computed as follows:

$$1.0 * n_f + 0.5 * n_p + 0.0 * n_n$$

Where n_f, n_p and n_n are the number of fluent, partially fluent and non-fluent translations, respectively. Adequacy values are computed in the same way. On the in-domain test set, both baselines achieve high adequacy and fluency, with the NMT baseline effectively matching the adequacy and fluency of the reference translations.

Regarding adequacy, the in-domain samples contain only a small number of translations with content unrelated to the source (1% to 2%). On out-of-domain data, on the other hand, both baselines produce a high number of inadequate translations: 57% (SMT) and 84% (NMT).



(a) in-domain (b) out-of-domain

Figure 2: Manual evaluation of adequacy and fluency for DE→EN. Legend: marker colors are different systems, marker types are different domains. SR=Subword Regularization, D=Distillation, R=Reconstruction

	in-domain	OOD average
Reference	2%	2%
NMT	2%	35%
SMT	1%	4%
NMT + Subword Regularization	1%	37%
NMT + Distillation	3%	33%
NMT + Reconstruction	1%	29%

Table 4: Percentage of translations judged to be hallucinations (both **not adequate** and at least **partially fluent**) in the manual evaluation.

These results suggest that the extremely low BLEU scores on these two test sets (see Table 2a) are in large part due to made up content in the translations.

Regarding fluency in out-of-domain settings, SMT and NMT baselines behave very differently: SMT translations are more adequate, while NMT translations are more fluent. This trend is most extreme in the *koran* domain, where only 2% of SMT translations are found to be fluent (compared to 36% for NMT).

Further analysis of both annotations shows that NMT translations found to be inadequate are not necessarily disfluent in out-of-domain settings. Table 4 shows that, on average, on out-of-domain data, 35% of NMT translations are both inadequate and fluent, while the same is only true for 4% of SMT translations. We refer to translations of this kind as **hallucinations**.

To summarize our analysis of baseline models, we find that the domain robustness of current NMT systems is still lacking and that inadequate, but fluent translations are a prominent issue. This motivates our choice of techniques to improve domain robustness.

4 Approaches to Improve Domain Robustness

We discuss approaches that can potentially remedy the problem of low domain robustness, and compare them in subsequent experiments.

4.1 Subword Regularization

Subword regularization (Kudo, 2018) is a form of data augmentation that, instead of applying a fixed subword segmentation like BPE, probabilistically samples a new subword segmentation

for each epoch. At test time, the model either uses 1-best segmentation, or translates the k-best segmentations and selects the highest-probability translation.

Kudo (2018) reports large improvements on low-resource and out-of-domain settings. In particular, improvements on in-house patent, web, and query test sets were in the range of 2–10 BLEU. In this work, we evaluate subword regularization on public datasets. We apply sampling at training time and translate 1-best segmented sentences at test time.

4.2 Defensive Distillation

We hypothesize that defensive distillation can be used to improve domain robustness. Defensive distillation exploits *knowledge distillation* to fend off adversarial attacks.

Knowledge distillation is a technique to derive models from existing ones, instead of training from scratch. The idea was introduced for simple image classification models by Ba and Caruana (2014) and Hinton et al. (2015). A first model (called the teacher) is trained in the usual fashion. Then, a second model (called the student) is trained using the predictions of the teacher model instead of the labels in the training data.

Typically, knowledge distillation is used to approach the performance of a complex teacher model (or ensemble of models) with a simpler student model. Another application is *defensive distillation* (e.g. Papernot et al. 2016), where the student shares the network architecture with the teacher, with the purpose not being model compression, but improving the model's generalization to samples outside of its training set, and specifically robustness against adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2014).

Defensive distillation has been shown to be effective at improving robustness to adversarial examples in image recognition. In this work, we apply it to NMT and test its effect on domain robustness, which Papernot et al. (2016) hint at but do not empirically test. As proposed by Kim and Rush (2016), we train the student model on teacher translations produced with beam search, instead of training on soft prediction labels.

4.3 Reconstruction

Reconstruction (Tu et al., 2017) is a change to the model architecture that addresses the problem of adequacy. The authors propose to extend encoder-decoder models with a *reconstructor* component that learns to reconstruct the source sentence from decoder states. The reconstructor has two uses: as a training objective, it forces the decoder representations to retain information that will be useful for reconstruction; during inference, it can provide scores that can be used for reranking.

However, we observed in initial experiments that reconstruction from hidden states can be too easy: the reconstruction loss on training batches diminishes very quickly, to the point of being insignificant. To prevent the model from simply reserving parts of the decoder hidden states to memorize the input sentence, we use reconstruction from actual translations instead of hidden states (Niu et al., 2019). Translations are produced with differentiable sampling via the Straight-Through Gumbel Softmax (Jang et al., 2017), which still allows joint optimization of translation and reconstruction. While Niu et al. (2019) implement reconstruction for recurrent architectures, we apply the technique to Transformers.

In order to avoid introducing any additional parameters for reconstruction, as recommended in Niu et al. (2019), we train a multilingual, bi-directional system with shared parameters as a further baseline. This bi-directional system is used to initialize the fine-tuning of reconstruction models. We empirically test whether our bilingual baseline and this multilingual baseline have comparable performance.

4.4 Neural Noisy Channel Reranking

Even though the methods we presented previously do lead to improved out-of-domain translation quality, the models still suffer from low adequacy. Also, our reconstruction models only perform reconstruction during training, and the reverse translation direction is not exploited, for instance by reranking translations (Tu et al., 2017).

We conjecture that this problem can be addressed with a neural noisy channel model (Li and Jurafsky, 2016). Standard NMT systems only model $p(y|x)$, which can lead to a "failure mode that can occur in conditional models in which inputs are explained away by highly predictive output prefixes" (Yu et al., 2017). Noisy channel models propose to also model $p(x|y)$ and $p(y)$ to alleviate this effect.

In practical terms, noisy channel models are implemented by modifying the core decoding algorithm, or simply as n-best list reranking. We adopt the latter, since n-best list reranking was shown to have equal or better performance than more computationally costly methods that score partial hypotheses during beam search (Yee et al., 2019).

5 Experimental Setup for Proposed Methods

This section describes how we preprocessed data and trained the models described in Section 4. Unless stated otherwise, the data is preprocessed in the same way as for the baselines (see Section 3.1).

Subword Regularization Models We integrate subword regularization in Sockeye, using the Python library provided by Kudo (2018)[4]. The training data is **not** segmented with BPE in this case. Instead, the training tool is given truecased data, and new segmentations are sampled before each training epoch. We use the following hyperparameters: we set the smoothing parameter α to 0.1 and use an n-best size of 64. For the validation and test data we use 1-best segmentation.

Defensive Distillation Models We use our baseline Transformer model as the teacher model. We translate the original training set with beam size 10. The student is trained on the translations of the teacher model using the same hyperparameters and being initialized with the parameters of the teacher model.

Reconstruction Models We implement differentiable sampling reconstruction for Transformer models in Sockeye and release the implementation.[5] We first train a multilingual Transformer model using the approach of Johnson et al. (2017).

After early stopping, we continue training with reconstruction as an additional loss. All hyperparameters remain the same, except for the new loss and a lower initial learning rate. For testing we select the model with the lowest validation perplexity. We use the reconstruction loss only for training, not for reranking.

Noisy Channel Reranking For each hypothesis, we store an n-best list of 50. We produce the following scores: $p(y|x)$ (usual translation score), $p(x|y)$ (translation score in reverse direction) and $p(y)$ (language model score in target language). $p(y|x)$ and $p(x|y)$ are computed with the same model since it is bi-directional.

In order to produce $p(y)$ scores we train a Transformer language model with fairseq (Ott et al., 2019) using standard settings. We impose a large penalty of -100 for hypotheses that contain subwords not found in the target side training data.

The final hypothesis score for reranking is computed as a weighted multiplication:

$$\text{score}(x, y) = p(y|x)^{\lambda_{tf}} * p(x|y)^{\lambda_{tb}} * p(y)^{\lambda_{lm}}$$

[4] https://github.com/google/sentencepiece

[5] https://github.com/ZurichNLP/sockeye/tree/domain-robustness

	DE→EN			DE→RM		
	λ_{tf}	λ_{tb}	λ_{lm}	λ_{tf}	λ_{tb}	λ_{lm}
(5) Multilingual	0.7	0.26	0.04	0.9	0.09	0.01
(6) Reconstruction	0.6	0.32	0.08	0.9	0.09	0.01
(7) Multilingual + SR	0.5	0.42	0.08	0.9	0.09	0.01
(8) Reconstruction + SR	0.5	0.46	0.04	0.9	0.09	0.01

Table 5: Best weights for noisy channel reranking found with grid search on in-domain development set. Row numbers correspond to the ones in Table 6. λ_{tf}=forward translation weight, λ_{tb}=backward translation weight, λ_{lm}= language model weight

The best weights are found with simple grid search on the in-domain development set, with $\lambda_{tf} \in [0.0, 0.1, ..., 1.0]$, $\lambda_{lm} \in [0.0, 0.01, ..., 0.1, 0.2, ..., 1.0]$, and $\lambda_{tb} = 1 - \lambda_{tf} - \lambda_{lm}$. The best weight combination is then used to compute scores and perform reranking for the test data of all domains. Table 5 lists optimal weights found for each model individually.

6 Results

We evaluate models in terms of BLEU on different domains (see Section 6.1) and also annotate a subset of translations manually for fluency and adequacy (see Section 6.2).

6.1 Automatic evaluation

Table 6 shows the results of our automatic evaluation. Overall, the proposed methods are able to outperform the SMT and NMT baselines but not in a consistent manner across domains or data conditions: an increase in in-domain BLEU does not always mean a similar increase on out-of-domain data. Similarly, an increase in BLEU in high-resource conditions (DE→EN) does not consistently lead to better results in low-resource conditions (DE→RM).

Subword regularization The results for subword regularization are mixed (see Row 3 in Table 6). For DE→EN, in-domain translation quality is comparable to the NMT baseline, while the average out-of-domain BLEU falls short of the NMT baseline (-0.5 BLEU). However, in the low-resource condition (DE→RM), subword regularization improves both in-domain and out-of-domain translation (+1.2 in both cases).

This result is surprising given the larger gains reported by Kudo (2018). To validate our implementation of subword regularization, we reproduce an experiment from Kudo (2018) with English-Vietnamese data from IWSLT 15 (see Table 7). With subword regularization we observe an improvement of 0.8 BLEU, which is lower than the 2 BLEU improvement reported by Kudo (2018), but we also note that our baseline model is stronger.

If subword regularization is combined with multilingual or reconstruction models (see Rows 7 and 8 in Table 6), we observe no improvements on in-domain test sets, but gains on 3 out of 4 out-of-domain data sets, indicating that subword regularization is in fact helpful for domain robustness.

Defensive Distillation Distilling the training data also leads to improvements in BLEU on out-of-domain text (see Row 4 in Table 6). The average gain is +1.4 for DE→EN, but only +0.4 for DE→RM. In-domain performance is either comparable or slightly worse (-0.4 BLEU for DE→EN) than the NMT baseline. The technique was originally shown to guard against adversarial attacks, where inputs are only infinitesimally different from training examples. Our results indicate that generalization to out-of-domain inputs – that are farther from the training data – is similarly improved.

	DE→EN		DE→RM	
	in-domain	average OOD	in-domain	average OOD
(1) SMT	58.4	11.8	45.2	15.5
(2) NMT	61.5	11.7	52.5	18.9
(3) NMT + SR	61.4	11.2	**53.7**	20.1
(4) NMT + D	61.1	13.1	52.5	19.3
(5) Multilingual	61.4	11.7	52.8	19.6
(6) Reconstruction	61.5	12.5	53.4	21.2
(7) Multilingual + SR	60.3	12.8	52.4	20.1
(8) Reconstruction + SR	60.3	**13.2**	52.4	20.3
(9) Multilingual + NC	62.7	11.8	53.1	21.4
(10) Reconstruction + NC	**62.8**	13.0	53.3	**21.6**
(11) Multilingual + SR + NC	60.7	12.3	53.1	21.4
(12) Reconstruction + SR + NC	60.8	13.1	52.4	20.7

Table 6: BLEU scores (higher is better) of all systems on test data. SR=Subword Regularization, D=Distillation, NC=Noisy Channel Model, average OOD=average BLEU score over out-of-domain test sets.

	Baseline (BPE)	Subword Regularization
Kudo (2018)	25.6	27.7 (+ 2.1)
Our results	28.3	29.1 (+ 0.8)

Table 7: Reproducing results from Kudo (2018) on IWSLT 15 English-Vietnamese data.

Source	- die Produktion in der Türkei entspricht 1,3 % der chinesischen Produktion;
Target	- Turkey's volume of production amounts to 1,3 % of Chinese production,
NMT Baseline	- the production in slkei is 1.3% of a Chinese **hamster ovary (CHO) cell**
Multilingual	- production in turkei is equivalent to 1.3% of Chinese **Hamster** production;
Reconstruction	- the production in thekei is equivalent to 1.3% of the Chinese production;
Reconstruction + NC	- production in the turkei equals 1.3% of the Chinese production;

Table 8: Example translations for DE→EN. Hallucinated parts are shown in **bold**.

Reconstruction Since reconstruction models are fine-tuned from multilingual models, we report scores for those multilingual models as well. Row 5 of Table 6 shows that our multilingual models perform equally well or better than the NMT baseline. As shown in Row 6 of Table 6, reconstruction outperforms the NMT baseline on out-of-domain data for both language pairs (+0.8 BLEU for DE→EN, +2.3 BLEU for DE→RM), while maintaining (DE→EN) or improving (+0.9 BLEU for DE→RM) in-domain BLEU.

Noisy Channel Reranking We evaluate the performance of noisy channel reranking in four different settings: applied to multilingual or reconstruction systems, both with and without subword regularization. The results are shown in Rows 9 to 12 of Table 6.

For DE→EN, reranking a reconstruction model achieves a good in-domain BLEU (+1.3 over the baseline), and slightly improves out-of-domain translation on average (+0.5 BLEU over reconstruction). For DE→RM in our low-resource setting, reranking with a noisy channel model improves the reconstruction model by +0.4 BLEU, producing the best result overall. The improvement on out-of-domain translation is much larger for the multilingual model (+1.8 over multilingual model without reranking). Combining reranked models (see Rows 11 and 12 in Table 6) with subword regularization does not lead to consistent improvements. Out-of-domain BLEU for DE→RM is slightly better compared to a subword regularization system without reranking (+0.4 BLEU), while all other scores are comparable or worse.

We believe that the success of reranking could be limited for two reasons. Firstly, model weights λ are optimized on in-domain data. Oracle experiments that optimize weights on out-of-domain data show that optimal model weighting differs greatly between domains and could further improve out-of-domain results by 0.5–1 BLEU on average. Secondly, reranking could be held back by the lack of diversity among hypotheses in n-best lists.

6.2 Manual evaluation

In a small manual evaluation, we analyze if methods that improve domain robustness in terms of BLEU also directly address the lack of adequacy in out-of-domain translation (see Section 3.2.1). The results are shown in Figure 2b. For anectodal examples of translations by several models, see Table 8.

Techniques with a bias for fluency We find that subword regularization increases fluency by several percentage points in some domains. However, it fails to improve adequacy which explains why it consequently fails to improve out-of-domain BLEU. Defensive distillation improves translations in a similar way, with a bias for fluency while not consistenly improving the adequacy of out-of-domain translations.

Reconstruction having a bias for adequacy We show that reconstruction is limiting hallucination on out-of-domain data, reducing the percentage of inadequate translations by 5 percentage points on average. We also note that there appears to be a tradeoff between adequacy and fluency: while reconstruction does improve out-of-domain adequacy, the improvement comes at the cost of lower fluency.

7 Discussion

Overall, we emphasize that SMT no longer appears to have higher domain robustness than NMT—a widely held belief given the findings of Koehn and Knowles (2017). However, SMT and NMT models achieve comparable domain robustness in different ways: SMT translations are more adequate (since SMT decoding implicitly enforces coverage), while NMT translations are more fluent (since an NMT decoder is simply a language model of the target language conditioned on source context). Therefore, evaluating domain robustness with automatic metrics only can be misleading and hide the fact that models have very different inductive biases, even if their BLEU score is similar.

Our experiments to reduce NMT hallucination and improve adequacy can be summarized as follows. Several methods lead to moderate improvements in translation quality, but only reconstruction combined with noisy channel reranking robustly increased BLEU across domains and data conditions (see Row 10 of Table 6). Other techniques are successful only in certain conditions. For instance, combining multilingual systems with Subword Regularization improves out-of-domain translation only, but not in-domain quality (see Rows 7–8 of Table 6). This suggests that techniques found to work well on in-domain test data cannot be assumed to have a similar effect on out-of-domain data without proper testing.

Our manual evaluation suggests that for NMT systems, an increase in out-of-domain BLEU also means that translations are slightly more adequate. However, NMT models still have a strong bias for fluency (all models under the diagonal in Figure 2b) and their adequacy falls short of the adequacy of SMT systems. Our most successful reconstruction model does indeed improve adequacy, but at the same time decreases fluency to a certain extent. We believe radically different approaches are needed to increase the coverage and adequacy of NMT translations without sacrificing their fluency.

8 Conclusions

Current NMT systems exhibit low domain robustness, i.e. they underperform if they are tested on a domain that differs strongly from the training domain. This is especially problematic in settings where explicit domain adaptation is impossible because the target domain is unknown, or because we are in a low-resource setting where training data is only available for limited domains. We find that hallucinated translations are a common problem for NMT models in out-of-domain settings, which partially explains their low domain robustness. Our results show that several methods yield improved generalization to out-of-domain data. We achieve an improvement in average out-of-domain BLEU of 1.5 (DE→EN) and 2.7 (DE→RM), as well as a reduction in hallucinated translations according to manual evaluation.

We analyzed the fluency and adequacy of translations manually, leading to a discussion of several pitfalls regarding the evaluation of NMT models in out-of-domain settings. Also, we believe that future research will need to address the lack of adequacy without losing fluency. We consider domain robustness an unsolved problem and encourage further research. For this purpose, we share data and code to serve as a baseline for future experiments.[6]

Acknowledgements

We are grateful to Manuela Weibel, Beni Ruef, Florentin Lutz and the *Lia Rumantscha* for providing Rumansh data, to Xing Niu for his help with reconstruction training, and to *Fremde Welten* for their help with manual annotations. MM and AR have received funding from the Swiss National Science Foundation (SNF, grant number 105212_169888). RS has received funding from the European Union's Horizon 2020 Research and Innovation Programme (ELITR, grant agreement 825460).

[6] https://github.com/ZurichNLP/domain-robustness

References

Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of EMNLP 2016*, pages 1557–1567.

Ba, J. and Caruana, R. (2014). Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662.

Belinkov, Y. and Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada.

Goodfellow, I., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples (2014). *arXiv preprint arXiv:1412.6572*.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2018). The Sockeye Neural Machine Translation Toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 200–207.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparametrization with gumbel-softmax. In *Proceedings International Conference on Learning Representations 2017*. OpenReviews.net.

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

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas.

Kobus, C., Crego, J., and Senellart, J. (2017). Domain control for neural machine translation. In *Proceedings of RANLP 2017*, pages 372–378. INCOMA Ltd.

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 ACL-2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

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.

Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of ACL 2018*, pages 66–75.

Läubli, S., Müller, M., Horat, B., and Volk, M. (2018). mtrain: A convenience tool for machine translation.

Lee, K., Firat, O., Agarwal, A., Fannjiang, C., and Sussillo, D. (2018). Hallucinations in neural machine translation.

Li, J. and Jurafsky, D. (2016). Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*.

Li, X., Michel, P., Anastasopoulos, A., Belinkov, Y., Durrani, N., Firat, O., Koehn, P., Neubig, G., Pino, J., and Sajjad, H. (2019). Findings of the First Shared Task on Machine Translation Robustness. In *WMT 2019*, pages 91–102, Florence, Italy.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of LREC 2016*, pages 923–929.

Luong, M.-T. and Manning, C. D. (2015). Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of IWSLT 2015*, Da Nang, Vietnam.

Nguyen, T. and Chiang, D. (2018). Improving Lexical Choice in Neural Machine Translation. In *Proceedings of NAACL-HLT 2018*, pages 334–343.

Niu, X., Xu, W., and Carpuat, M. (2019). Bi-directional differentiable input reconstruction for low-resource neural machine translation. In *Proceedings of NAACL-HLT 2019*, pages 442–448.

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 NAACL-HLT 2019: Demonstrations*.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE.

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 on Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

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.

Scherrer, Y. and Cartoni, B. (2012). The Trilingual ALLEGRA Corpus: Presentation and Possible Use for Lexicon Induction. In *Proceedings of LREC-2012*, pages 2890–2896, Istanbul, Turkey.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of ACL 2016*, pages 86–96, Berlin, Germany.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL 2016*, pages 1715–1725, Berlin, Germany.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tu, Z., Liu, Y., Shang, L., Liu, X., and Li, H. (2017). Neural Machine Translation with Reconstruction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3097–3103.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Yee, K., Dauphin, Y., and Auli, M. (2019). Simple and effective noisy channel modeling for neural machine translation. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5696–5701, Hong Kong, China.

Yu, L., Blunsom, P., Dyer, C., Grefenstette, E., and Kociský, T. (2017). The neural noisy channel. In *5th International Conference on Learning Representations, ICLR 2017*, Toulon, France.