Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis
Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez
Dep. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03690 Sant Vicent del Raspeig (Spain)
{vmsanchez,mespla,japerez,fsanchez}@dlsi.ua.es

Abstract

In the context of neural machine translation, data augmentation (DA) techniques may be used for generating additional training samples when the available parallel data are scarce. Many DA approaches aim at expanding the support of the empirical data distribution by generating new sentence pairs that contain infrequent words, thus making it closer to the true data distribution of parallel sentences. In this paper, we propose to follow a completely different approach and present a multi-task DA approach in which we generate new sentence pairs with transformations, such as reversing the order of the target sentence, which produce unfluent target sentences. During training, these augmented sentences are used as auxiliary tasks in a multi-task framework with the aim of providing new contexts where the target prefix is not informative enough to predict the next word. This strengthens the encoder and forces the decoder to pay more attention to the source representations of the encoder. Experiments carried out on six low-resource translation tasks show consistent improvements over the baseline and over DA methods aiming at extending the support of the empirical data distribution. The systems trained with our approach rely more on the source tokens, are more robust against domain shift and suffer less hallucinations.

1 Introduction

In order to train reliable neural machine translation (NMT) systems, large amounts of parallel sentences —sentence pairs in two languages that are mutual translations— are needed, which constitutes a critical barrier for low-resource language pairs. This problem has been addressed through different approaches, such as transfer-learning from high-resource language pairs (Kocmi and Bojar, 2018), using linguistic annotations (Sennrich and Haddow, 2016), training multilingual systems (Johnson et al., 2017) and applying data augmentation strategies (Li et al., 2019; Feng et al., 2021), i.e., artificially generating additional parallel sentences.

Data augmentation (DA) is formalised by many authors as a solution to a data distribution mismatch problem (Wang et al., 2018; Wei et al., 2020). The data distribution of the sentence pairs observed in the training corpus, $\hat{p}(\mathbf{x}, \mathbf{y})$, differs from the true data distribution, $p(\mathbf{x}, \mathbf{y})$. Hence, the system should be trained on a training set that follows $q(\mathbf{x}, \mathbf{y})$, an augmented version of $\hat{p}(\mathbf{x}, \mathbf{y})$ with a wider support. In this way, the trained system is less likely to face totally out-of-distribution data when translating.

In this paper, we propose a completely different framework for DA in which we generate additional parallel sentences which, despite being completely unlikely under the data distribution, systematically improve the quality of the resulting NMT system. Inspired by one-to-many multilingual NMT, where richer encoder representations are obtained (Dong et al., 2015), we propose a set of simple DA strategies to produce synthetic target sentences aimed at strengthening the encoder. These strategies expose the network during training to new situations where the target-language context is not sufficient to achieve a low loss, and the burden is passed on to the encoder. Recent findings by Voita et al. (2021) further motivate our approach: they claim that the influence of source tokens in the output predictions of an NMT system decreases as decoding advances. Moreover, to avoid harmful interferences by the out-of-distribution target data generated, we follow a simple multi-task learning (MTL) approach that does not require changes to the model architecture. We call the proposed framework **multi-task learning data augmentation** (MTL DA) to stress the fact that the augmented data, which do not follow the distribution of parallel sentences in the training corpus, constitute different auxiliary tasks that nevertheless produce a positive transfer to the main task.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8502–8516
November 7–11, 2021.

Our framework does not require elaborated preprocessing steps, training additional systems, or data besides the available training parallel corpora. Experiments with six low-resource translation tasks show that it systematically outperforms two powerful methods aiming at extending the support of the empirical data distribution —RAML (Norouzi et al., 2016) and SwitchOut (Wang et al., 2018)— and that it can be combined with synthetic corpora generated through back-translation (Sennrich et al., 2016b) to get further improvements.

In the context of explainable deep learning models, we perform an analysis of the relevance of the encoder and decoder representations in the NMT system output, which shows that, thanks to the auxiliary tasks, MTL DA increases the contribution of the source tokens to the decisions made by the NMT system. Moreover, systems trained with MTL DA are much more robust against domain shift and produce less hallucinations (Wang and Sennrich, 2020).

The remainder of the paper is organised as follows. Next section briefly describes our MTL DA approach and the different auxiliary tasks we evaluated. After that, Sec. 3 describes the experimental settings, whereas Sec. 4 provides and discusses the results obtained. Sec. 5 then presents an analysis of the changes in the transformer dynamics induced by our auxiliary tasks as a way of explaining the improvement in translation quality. The paper ends with a review of the most relevant works in the area of DA for NMT in Sec. 6, followed by some concluding remarks in Sec. 7.

2 Multi-task learning approach and auxiliary tasks

We propose a simple MTL approach that consists of using a vanilla NMT system —in our experiments, it is a transformer system as defined by Vaswani et al. (2017)— where all (main and auxiliary) tasks share the encoder and the decoder. In order to avoid harmful interferences by the out-of-distribution target data generated for the auxiliary tasks, we add a task-specific artificial token to the source sentence to constrain the kind of output to be produced (Sennrich et al., 2016a; Johnson et al., 2017), much like in multilingual NMT. For each auxiliary task, we append a synthetic corpus of the same size to the original training data, which is obtained by applying a transformation to each original pair of sentences. In almost all the tasks, the source sentence is left unchanged while the target sentence is substantially modified.

What follows is a brief explanation of the transformations we have tested and their expected effect on the training dynamics of the encoder. Some of them have been previously applied in DA, but never in an MTL set-up such as the one we are presenting. Some transformations are controlled by a hyperparameter α that determines the proportion of target words affected by the transformation. In what follows, t denotes the amount of words in the original target sentence. Table 1 provides an example of the effect of the different transformations on a single sentence pair.

swap: Pairs of random target words are swapped until only $(1-\alpha)\cdot t$ words remain in their original position. This task (Artetxe et al., 2018; Lample et al., 2018) tries to force the system to trust less the target prefix when generating a new word.

token: $\alpha \cdot t$ random target words are replaced by a special (UNK) token (Xie et al., 2017). Again, when generating a new word, the target prefix should become less informative and force the system to pay more attention to the encoder. This is the effect envisaged by word dropout when preventing posterior collapse in variational autoencoders (Bowman et al., 2016).

source: The target sentence becomes a copy of the source sentence. In this way, the most efficient way of emitting the right output is checking the encoder representation to copy from the source. Some authors have identified such training instances as harmful for NMT (Ott et al., 2018; Khayrallah and Koehn, 2018), and only copying in the inverse direction has been proved to be useful (Currey et al., 2017). However, the MTL framework may allow us to leverage such synthetic training data.

reverse: The order of the words in the target sentence is reversed. Voita et al. (2021) suggest that the influence of the encoder decreases along the target sentence; therefore, by reversing the order we expect the system to learn to use more information from the encoder when generating words that usually appear near the end of the sentence.

mono: Target words are reordered so as to make the alignment between source and target words monotonous. This transformation uses one-to-many word alignments and is inspired by the con-

Task	Lang.	Synthetic training sample
original training sample	source	Es gibt andere Möglichkeiten , die Pyramide zu durchbrechen .
	target	There 's other ways of breaking the pyramid .
swap	target	There . other ways of breaking pyramid 's the
token	target	There 's other UNK of UNK UNK UNK .
source	target	Es gibt andere Möglichkeiten , die Pyramide zu durchbrechen .
reverse	target	. pyramid the breaking of ways other 's There
mono	target	's There other ways the pyramid of breaking .
replace	source	Es gibt aufzurüsten kalt , Schach Spezialwissen zu durchbrechen .
	target	There 's arming cold of breaking chess specialties .

Table 1: A German–English, word-aligned training sample (first row) and the result of applying the transformations described in Sec. 2 using $\alpha = 0.5$ for those transformations controlled by this hyperparameter. Words modified by each transformation are coloured; for *swap* and *replace*, a different colour identifies each pair of words that are either swapped or replaced together, respectively.

cept of *biwords* introduced by Sánchez-Martínez et al. (2012) for the compression of parallel corpora. By making the alignment between source and target words monotonous, the target sentences become less fluent, so we expect the system to pay more attention to the encoder.

replace: $\alpha \cdot t$ source–target aligned words are selected at random and replaced by random entries in a bilingual lexicon obtained from the training corpus; to this end, one-to-one word alignments are used.[1] This transformation is likely to introduce words that are difficult to produce by relying only on the target language prefix, thus forcing the system to pay attention to the source words. Fadaee et al. (2017) followed a similar approach; however, they constrained the replacements to produce only fluent target sentences.

3 Experimental settings

We have conducted experiments for the translation from English to German, Hebrew and Vietnamese, and for the translation in the reverse direction, using corpora commonly used for evaluating DA techniques in low-resource scenarios. We evaluated the effect of using each of the MTL DA auxiliary tasks, as well as the combination of the best performing ones. We also evaluated two strong DA methods that aim at extending the support of the empirical data distribution by replacing some words by random samples from the vocabulary: SwitchOut (applied on the source side), RAML (applied on the target side), and the combination of both.

[1] If the number of aligned words is below $\alpha \cdot t$, all available alignments are used.

Datasets. Following Gao et al. (2019) and Guo et al. (2020), for English–German and English–Hebrew we used the training data (speeches of TED and TEDx talks) of the IWSLT 2014 text translation track (Cettolo et al., 2014);[2] for development and testing we used the *tst2013* and *tst2014* datasets, respectively. Like Wang et al. (2018), for English–Vietnamese we used the training data (also TED talks) of the IWSLT 2015 text translation track (Cettolo et al., 2015);[3] datasets *tst2012* and *tst2013* were used, respectively, for development and testing.

To evaluate the impact of MTL DA when it is combined with synthetic data obtained through back-translation (Sennrich et al., 2016b) —that is, by translating a target-language monolingual corpus into the source language—, we collected additional English monolingual data, back-translated it into German, Hebrew and Vietnamese, and added the resulting synthetic corpus to the training data listed above. The monolingual data used consists of all the available monolingual English sentences in the IWSLT 2018 shared task on low-resource MT of TED talks that were not present in the parallel training data described above. Table 2 provides the amount of sentences and tokens in the training corpora used in our experiments.

In order to study the domain robustness of our MTL DA approach, we also evaluated the systems trained on some out-of-domain test sets. We chose the IT, law and medical test sets released by Müller

[2] https://sites.google.com/site/iwsltevaluation2014/data-provided

[3] https://wit3.fbk.eu/2015-01

Pair	# sent.	# left tok.	# right tok.
IWSLT parallel data only			
en–de	174,443	3,575,407	3,353,855
en–he	187,817	3,862,985	2,958,136
en–vi	133,317	2,965,962	3,361,789
IWSLT parallel data + back-translated data			
en–de	269,213	5,843,264	5,537,986
en–he	282,587	6,130,842	4,728,840
en–vi	228,087	5,413,428	6,232,006

Table 2: Number of sentences and tokens in the training corpora used in our experiments.

et al. (2020)[4] and also used by Wang and Sennrich (2020) for English–German.

Corpora were tokenised and truecased with the Moses scripts;[5] then, sentences with more than 100 or less than 5 tokens were removed from the training corpora. Afterwards, byte-pair encoding (BPE) with 10,000 merge operations (Sennrich et al., 2016c) was applied on the concatenation of the source and target sides of the training corpora to obtain the vocabulary. Finally, those sentence pairs in the training corpora with more than 100 BPE tokens were removed.

One-to-many word alignments in both translation directions were obtained using `mgiza++` (Gao and Vogel, 2008; Och and Ney, 2003).[6] Source-to-target word alignments were used for the *mono* transformations; the one-to-one word alignments required by the *replace* transformation were obtained by computing the intersection between the one-to-many word alignments in both translation directions. The bilingual lexicon for the *replace* transformation was built by associating to each source word the target word it is most frequently aligned with in the one-to-one word alignments.

Training. Our neural model is a transformer-based model as defined by Vaswani et al. (2017), with the exception of the amount of *warmup_steps*, which was set to 8,000. All the experiments were carried out on a single GPU with mini-batches made of 4,000 tokens. Validation was done every 1,000 updates, and the patience based on the BLEU score on the development set was set to 6 validation cycles; we then kept the intermediate model performing best on the development set. We trained the systems with the `fairseq` toolkit (Ott et al., 2019). For RAML and SwitchOut, we integrated into `fairseq` the sampling function released by Wang et al. (2018).[7]

Systems trained with MTL DA were fine-tuned on the main (translation) task after being trained on the combination of the main and auxiliary tasks. When combining different auxiliary tasks, a different special token was used for each one.

DA hyperparameters. The proportion of words affected by the *swap*, *token* and *replace* transformations is controlled by a hyperparameter α, whereas RAML and SwitchOut are governed by a temperature τ. For each language pair, we explored values of α in $[0.1, 0.9]$ at intervals of 0.1, and values of τ around the best values reported by Wang et al. (2018).[8] The results reported are those obtained with the model that maximizes BLEU on the development set. The best hyperparameters obtained for the experiments with the IWSLT parallel data were reused for the experiments with the training set extended with back-translated data.

4 Results and discussion

IWSLT parallel data. Table 3 reports the mean and standard deviation of the translation performance, measured in terms of BLEU (Papineni et al., 2002),[9] of three different executions for each of the systems trained on the IWSLT parallel data. chrF++ scores (Popović, 2017), that show the same trend, are available in Appendix B. The results show that our MTL DA approach consistently outperforms the baseline system in all language pairs and translation directions. In general, the auxiliary tasks *reverse* (translation into the target language but in the reverse order) and *replace* (random replacement of target words and the source words they are aligned with) are the best performing ones. *swap* (random swapping of words) and *source* (copying the source sentence) often perform worse than the former tasks, which suggests that a non-systematic word order or a completely different vocabulary in the target could negatively influence the main task.

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

[5] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

[6] https://github.com/moses-smt/mgiza

[7] Code available at https://github.com/transducens/mtl-da-emnlp.

[8] $\tau^{-1} \in \{0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 1.0, 1.1, 1.2, 1.3\}$

[9] `sacrebleu` (Post, 2018) version string: `BLEU+case.mixed+lang.vi-en+numrefs.1+smooth.exp+tok.13a+version.1.5.0`

IWSLT parallel data only

Task	en-de	de-en	en-he	he-en	en-vi	vi-en
baseline	24.7 ± 0.2	30.0 ± 0.1	21.5 ± 0.3	32.4 ± 0.1	28.9 ± 0.1	27.5 ± 0.4
SwitchOut	25.3 ± 0.2	30.1 ± 0.2	21.6 ± 0.6	32.1 ± 0.4	28.5 ± 0.2	27.3 ± 0.6
RAML	25.4 ± 0.2	30.3 ± 0.1	21.9 ± 0.1	32.1 ± 0.1	28.6 ± 0.5	27.3 ± 0.5
SwitchOut+RAML	$\mathbf{25.7 \pm 0.4}$	30.3 ± 0.5	22.1 ± 0.4	32.1 ± 0.4	29.1 ± 0.4	27.5 ± 0.3
swap	25.1 ± 0.2	30.3 ± 0.1	22.1 ± 0.4	32.5 ± 0.6	28.8 ± 0.2	28.3 ± 0.6
token	25.4 ± 0.2	30.0 ± 0.3	21.5 ± 0.2	32.4 ± 0.8	29.3 ± 0.3	28.2 ± 0.3
source	25.3 ± 0.1	30.1 ± 0.4	21.5 ± 0.3	32.7 ± 0.2	28.9 ± 0.3	27.6 ± 0.2
reverse	$\mathbf{26.1 \pm 0.3}$	30.2 ± 0.1	22.4 ± 0.2	$\mathbf{33.4 \pm 0.3}$	29.4 ± 0.3	28.2 ± 0.4
mono	25.7 ± 0.1	30.4 ± 0.2	22.0 ± 0.1	32.5 ± 0.6	29.3 ± 0.4	27.7 ± 0.4
replace	$\mathbf{25.8 \pm 0.3}$	30.7 ± 0.2	22.5 ± 0.2	$\mathbf{33.5 \pm 0.3}$	29.5 ± 0.3	28.3 ± 0.9
reverse+replace	$\mathbf{26.3 \pm 0.1}$	$\mathbf{31.1 \pm 0.3}$	$\mathbf{22.9 \pm 0.2}$	$\mathbf{33.9 \pm 0.2}$	$\mathbf{30.1 \pm 0.5}$	28.8 ± 0.2
reverse+mono+replace	$\mathbf{26.4 \pm 0.6}$	$\mathbf{31.4 \pm 0.3}$	$\mathbf{23.2 \pm 0.3}$	$\mathbf{33.9 \pm 0.5}$	$\mathbf{30.5 \pm 0.2}$	$\mathbf{29.4 \pm 0.3}$

Table 3: Mean and standard deviation of the BLEU scores obtained when translating in-domain test sets with the baseline system, three other reference systems, and our MTL DA approach, using different auxiliary tasks and combinations of them. Systems were trained only on IWSLT parallel data. The best results for each language pair, and those falling within one standard deviation from them, are highlighted in bold.

IWSLT parallel data + back-translated data

Task	de-en	he-en	vi-en
baseline	31.3 ± 0.5	34.5 ± 0.1	29.3 ± 0.3
SwitchOut+RAML	$\mathbf{31.7 \pm 0.8}$	34.1 ± 0.7	$\mathbf{29.9 \pm 0.5}$
reverse+mono+replace	$\mathbf{32.3 \pm 0.1}$	$\mathbf{35.3 \pm 0.3}$	$\mathbf{30.4 \pm 0.7}$

Table 4: Mean and standard deviation of the BLEU scores obtained when translating in-domain test sets with the baseline system, the combination of SwitchOut and RAML, and the best combination of auxiliary task in our MTL DA approach. Systems were trained on a combination of parallel and back-translated data. The best results for each language pair, and those falling within one standard deviation from them, are highlighted in bold.

Interestingly, using the three best auxiliary tasks together further improves the performance, achieving the best results in all translation tasks with BLEU scores between 1.1 and 1.9 points over the baseline.[10] This suggests that different auxiliary tasks affect the encoder in different ways and are somehow complementary.

A comparison of our MTL DA approach with RAML, SwitchOut and their combination (SwitchOut+RAML) shows that our approach, being much simpler in nature, also outperforms them.

Back-translated data. Table 4 shows the results obtained with the training set extended with back-translated data. In these experiments, we only evaluated MTL DA with the best performing combination of auxiliary tasks. As for SwitchOut and RAML, we only evaluated their combination, which according to Table 3 performs better than any method in isolation. Although the differences are slightly smaller, we can observe the same trend in the results: MTL DA still outperforms the baseline and the combination of SwitchOut and RAML. In addition, the results show that MTL DA and back-translation are two complementary DA approaches.

Domain robustness. Concerning the out-of-domain evaluation, Table 5 shows the BLEU scores obtained by each system; chrF++ scores show the same trend and are provided in Appendix B. We restricted the MTL DA evaluation to the *reverse* task and the combination of the best three auxiliary tasks, and only report the results obtained with systems trained on the IWSLT parallel data.

As can be seen, MTL DA outperforms the baseline system and RAML/SwitchOut. Even the *reverse* auxiliary task, which does not modify the vocabulary of the target sentence in any way and does not add infrequent words to the training corpus, enhances the domain robustness of the system.

[10] Even though MTL DA is intended for low-resource language pairs, we conducted preliminary experiments on large training data using the English–German WMT 2014 dataset (Gao et al., 2019), which contains around 4.5M training parallel sentences. The results show a gain of around 1 BLEU point over the baseline for German–English and the same performance for English–German. In any case, the performance of MTL DA on large data sets remains to be studied.

Domain	IT		Law		Medical	
Direction	**en-de**	**de-en**	**en-de**	**de-en**	**en-de**	**de-en**
baseline	3.0 ± 0.3	6.2 ± 1.9	6.0 ± 0.7	8.1 ± 0.8	9.5 ± 0.6	10.7 ± 1.5
SwitchOut	5.1 ± 0.8	5.3 ± 2.5	7.6 ± 0.2	7.8 ± 0.5	11.8 ± 0.2	10.1 ± 1.2
RAML	5.0 ± 1.4	6.3 ± 3.1	7.8 ± 0.3	$\mathbf{8.5 \pm 0.7}$	11.7 ± 1.0	11.6 ± 1.6
SwitchOut+RAML	8.0 ± 0.5	6.8 ± 0.7	7.8 ± 0.2	7.5 ± 1.2	12.5 ± 1.1	10.1 ± 1.0
reverse	10.2 ± 0.6	$\mathbf{10.7 \pm 0.6}$	7.6 ± 0.4	8.0 ± 0.6	12.8 ± 0.4	12.1 ± 0.5
reverse+mono+replace	$\mathbf{13.2 \pm 1.5}$	$\mathbf{11.3 \pm 0.9}$	$\mathbf{10.2 \pm 0.4}$	$\mathbf{9.9 \pm 0.8}$	$\mathbf{15.9 \pm 1.0}$	$\mathbf{14.3 \pm 0.8}$

Table 5: Mean and standard deviation of the BLEU scores obtained when translating out-of-domain texts in the IT, law and medical domains. The best results for each language pair, and those falling within one standard deviation from them, are highlighted in bold.

5 Explainability

To confirm that the systematic improvements in translation quality and enhanced domain robustness are related to the encoder being exposed during training to more situations where a good source representation is crucial, we carried out an analysis of the relative source and target contributions to the generation decisions of the NMT system. According to Voita et al. (2021), systems trained with more data tend to rely more on source information; we expect MTL DA to produce the same effect.

Another aspect that will account for the positive impact of MTL DA in the system's encoder is the generation of *hallucinations* (Lee et al., 2018): completely inadequate translations that usually occur under domain shift (Müller et al., 2020), due to the system relying too much on the target context (Voita et al., 2021). We expect systems trained with MTL DA to produce less hallucinations. To validate this last hypothesis, we carried out an hallucination analysis on the results of the domain shift experiments.

Relative source and target contributions. We used *layer-wise relevance propagation* (LRP), as adapted to transformers by Voita et al. (2021), to compute the relative contribution of source and target tokens to each prediction made by the system. LRP allows us to compute $R_t(x_i)$ and $R_t(y_j)$, the relative contribution of source token x_i and target token y_j, respectively, to the prediction emitted by the network at time t. The total relevance at each time step is 1, i.e. for all time steps t the following equation holds, where $R_t(\mathbf{x})$ and $R_t(\mathbf{y})$ stand for the total contribution of the source tokens and target tokens, respectively:

$$\sum_i R_t(x_i) + \sum_j R_t(y_j) = R_t(\mathbf{x}) + R_t(\mathbf{y}) = 1$$

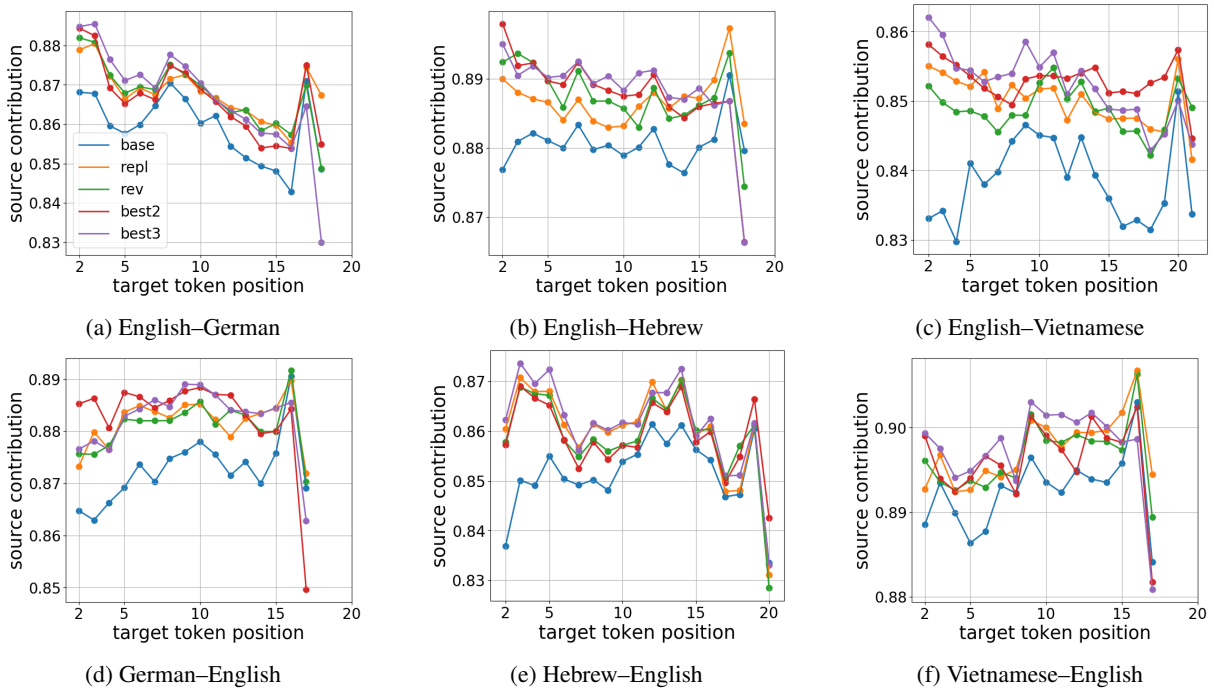
To have reliable comparisons, we also follow Voita et al. (2021) and evaluate the relative source and target contributions on a subset of sentences from a held-out corpus with the same source length and the same target length. In this way, we can fairly compare different strategies, since we teacher-force the reference translations when computing LRP so as to obtain translations with exactly the same length. The held-out corpus used is the concatenation of all the development corpora released for the corresponding IWSLT task,[11] while the subset chosen is the largest set of parallel sentences with the same source length and the same target length, as long as there are at least 16 tokens in each side.[12] To compute the relative contributions, we retrained the baseline and the MTL DA systems featuring the best performing auxiliary tasks on the IWSLT parallel data with the toolkit released by Voita et al. (2021).[13] Figure 1 depicts the total source contribution at each generation time step t for the different translation tasks and systems. We skip the first time step, when there is no target prefix available, and show the source contribution for the EOS token too. For MTL DA models, the source contains the special token corresponding to the main task.

For the translation tasks with English as source language, we can observe the same trends as Voita et al. (2021): source influence decreases as decoding progresses. There is also a peak in the penulti-

[11] English–German: dev2010, dev2012, tst2010, tst2011, tst2012, tst2013, tst2014; English–Hebrew: dev2010, tst2010, tst2011, tst2012, tst2013, tst2014; English–Vietnamese: dev2010, tst2010, tst2011, tst2012, tst2013.

[12] It contains 48 en–de sentences (768 en tokens and 816 de tokens), 28 en–he sentences (448 en tokens and 560 he tokens), and 28 en–vi sentences (448 en tokens and 560 vi tokens). Similar trends in source contribution were observed when selecting a larger subset with shorter sentences.

[13] https://github.com/lena-voita/the-story-of-heads



(a) English–German (b) English–Hebrew (c) English–Vietnamese

(d) German–English (e) Hebrew–English (f) Vietnamese–English

Figure 1: For each translation task, total contribution of the source tokens $R_t(\mathbf{x})$ to the production of the target tokens. *best2* stands for the combination of the two best auxiliary tasks within our MTL DA framework: *reverse* and *replace*. *best3* stands for the combination of the three best ones: *reverse*, *replace* and *mono*.

mate token, which may express that the decoder is checking whether it has translated all the content from the source sentence before emitting the full stop at the end of the sentence. When English is the target language, plots are flatter: source influence does not decrease as decoder advances. The fact that English grammar is simpler, lacking gender and case agreements, could explain that the decoder needs to check previous tokens less.

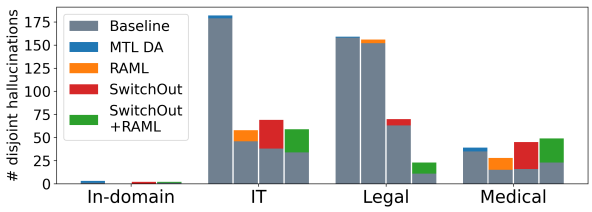
These results confirm the utility of MTL DA: the baseline system is systematically the one where the source has the smallest influence, and auxiliary tasks increase source influence in all translation tasks. Differences are larger at the beginning of decoding, but remain throughout the sentence. MTL DA achieves, only with artificially augmented data, an increase in source influence comparable to that reported by Voita et al. (2021, Fig. 6) when the size of genuine parallel data increases.

Finally, no consistent differences in source influence could be found between the *reverse* and *replace* auxiliary tasks. The systems combining multiple auxiliary tasks, however, are consistently the ones with the highest source influence, thus confirming the complementarity of the tasks.

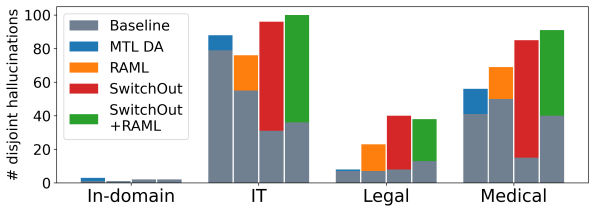
Hallucinations. To estimate the number of hallucinations produced by the systems evaluated, we follow the procedure proposed by Lee et al. (2018) and used by Raunak et al. (2021). Although their interest was in detecting those sentences that induced the generation of hallucinations after introducing spurious tokens in the input, we adapted it to automatically measure the number of input sentences in a test set for which the corresponding output seems to be an hallucination. To this end, we use an adjusted version of BLEU which only takes into account the precision of unigrams and bigrams with weights 0.8 and 0.2, respectively, as proposed by Lee et al. (2018). If the sentence-level adjusted BLEU of the lowercased emitted translation is below a certain threshold (10 in our experiments), it is taken as a sign of hallucination.

We evaluate the tendency to produce hallucination of the baseline as compared to our MTL DA approach combining the auxiliary tasks *reverse*, *mono* and *replace*, and to the SwitchOut, RAML and SwitchOut+RAML systems. Sentences whose translations cannot be regarded as hallucinations are not relevant to our study, neither are those for which the baseline and the system to which it is compared to both hallucinate. We therefore count the number of sentences that induce an hallucination on one of the systems but not on the other.[14]

[14] As an example, the evaluation corpus contains the pair



(a) English–German



(b) German–English

Figure 2: Number of disjoint hallucinations of the baseline system (in grey) and the systems trained with different DA methods. Our MTL DA approach (in blue) corresponds to the system labeled as *reverse+mono+replace* in Table 3.

We consider that the other system is not hallucinating if its adjusted BLEU is at least 20 BLEU points higher. Fig. 2 represents these data for the English–German translation tasks for the same domains (in-domain, IT, legal and medical) and corpora used for the domain robustness evaluation reported in Sec. 4. The grey bar represents the number of hallucinations of the baseline that are not labeled as such by the corresponding DA system; consequently, higher values are better. Additionally, the color bars represent the number of hallucinations of the DA system that cannot be labeled as such in the baseline system (shorter bars are better).

As can be seen, disjoint hallucinations barely happen with in-domain data, but they can be easily found when considering out-of-domain data. In every domain and in both translation directions, the blue bar (the one representing our MTL DA method) is clearly the shortest one in almost all cases, while at the same time the grey bar is the largest one. This is a clear sign of reduction of hallucinations in the systems trained with MTL DA. Appendix C presents supplementary details on the BLEU thresholds and the smoothing technique considered when computing the scores.

("To Select an Object", "Objekt auswählen"). The German–English baseline produces "Ein Objekt auszuwählen, um ein Objekt auszuwählen." which shows an hallucination in the form of a repetition, whereas the MTL DA method gives a better "Um ein Objekt auszuwählen.".

6 Related work

The back-translation (Sennrich et al., 2016b) approach for leveraging additional target monolingual data to produce additional training samples is probably the most popular DA approach for NMT. The set of related approaches covered in this section, however, mainly focus on methods that, as MTL DA, do not require additional resources besides the training parallel corpus.

Li et al. (2019) evaluate back- and forward-translation in such a setting. They train forward and backward NMT systems on the available parallel data and use them to produce new synthetic samples by translating either the target side (Sennrich et al., 2016b) or the source side (Zhang and Zong, 2016) of the original training corpus.

The approaches we have evaluated in our experiments, RAML (Norouzi et al., 2016) and its extension to the source language, SwitchOut (Wang et al., 2018), aim at extending the support of the empirical data distribution and keeping it smooth (similar sentence pairs have similar probabilities). To that end, they replace words with other words sampled from a uniform distribution over the vocabulary, which, in practice, results in infrequent words being overrepresented. Guo et al. (2020) presented a related approach to encourage compositional behaviour: replaced words are selected from another sentence and not from the vocabulary.

Some of our auxiliary tasks have already been used for DA, but mostly on the source side and rarely in an MTL framework. Replacing tokens with placeholders (as we do in *token*) has already been applied by Zhang et al. (2020) to the source language, in combination with auxiliary tasks involving detecting replaced and dropped tokens. Xie et al. (2017) also evaluate the impact of replacements on the target data, but do not follow an MTL approach. Word dropout (Sennrich et al., 2016b; Gal and Ghahramani, 2016; Shen et al., 2020) can also be considered a related approach.

Regarding changes to word order, in addition to the proposals by Artetxe et al. (2018) and Lample et al. (2018), it is worth highlighting the strategy proposed by Zhang et al. (2019) who apply a self-translation approach using a right-to-left decoder. Unlike our MTL DA framework, they need to generate translations from the model during training and adjust multiple terms in the training loss.

There are more DA approaches based on replacing words which are worth mentioning. Xie et al.

(2017) randomly replace words in the source side of the training samples. Gao et al. (2019) replace words selected at random with *soft words* whose representations are obtained from the probability distribution provided by a language model. Fadaee et al. (2017) replace a number of words in their training samples by infrequent words in order to improve the performance of the NMT model when dealing with them at translation time. Words to be replaced are identified using a large source language model. Once the source words to be replaced are identified, a word-alignment model and a probabilistic dictionary are used to also replace the corresponding counterpart by the most probable translation of the new source word. In our MTL DA framework, the *replace* transformation, which is similar to Fadaee et al. (2017)'s work, does not require any language model.

Regarding back-translation, Edunov et al. (2018) apply several simple transformations (word deletion, replacement, swapping) to back-translated data reporting a noticeable improvement. In relation with the special token we use to prevent negative transfer between tasks, Caswell et al. (2019) propose a similar strategy to identify synthetic samples when combining actual parallel data and back-translated data for training. Yang et al. (2019) extends this work by including forward-translated data for training using two different special tokens to distinguish the two types of synthetic data.

7 Concluding remarks

In this paper, we have presented a multi-task learning approach for data augmentation (MTL DA) in NMT. We deviate from common approaches that aim at extending the support of the empirical data distribution by generating new samples that are likely under such distribution. We propose instead to carry out DA in a MTL manner, by artificially generating new sentence pairs with aggressive transformations, such as reversing the order of the target sentence, which may make the target sentence completely unfluent. Translating into these augmented sentences constitute new tasks that provide new contexts during training where the target prefix is not informative enough to predict the next word, thus strengthening the encoder and forcing the system to rely more on it.

Experiments carried out on six low-resource translation tasks that usually serve as benchmark for DA show consistent improvements over a baseline system (on average around 1.6 BLEU points) and over strong DA methods that aim at extending the support of the empirical data distribution without MTL. Moreover, additional analyses show that the systems trained with MTL DA rely more on the source tokens, are more robust against domain shift and suffer less hallucinations.

MTL DA is agnostic to the NMT model architecture and does not require elaborated preprocessing steps, training additional systems, or data besides the available training parallel corpora. Furthermore, it could be combined with existing DA methods, in addition to back-translation, specially those that operate on the source side (Wang et al., 2018; Gao et al., 2019), since our transformations mainly address the target.

We expect this strategy to inspire the implementation of new auxiliary tasks to be used for MTL DA, specially those aimed at improving the training dynamics of the system. We believe that further improvements could be obtained by following more elaborated strategies for multi-task learning, such as changing the proportion of data for the different tasks, evaluating different ways of parameter sharing between the different tasks (e.g. sharing the encoder but not the decoder), and using other training schedules (Chen et al., 2018). Finally, we conclude that making the encoder representation essential to minimize the loss during training should be embraced as a potential way of boosting NMT quality.

Acknowledgments

We thank Wilker Aziz, a collaboration with whom served as an inspiration for this work, and Ivan Titov, who gave us valuable hints about the role of encoder representations. Work funded by the European Union's Horizon 2020 research and innovation programme under grant agreement number 825299, project Global Under-Resourced Media Translation (GoURMET); and by Generalitat Valenciana through project GV/2021/064. The computational resources used for the experiments were funded by the European Regional Development Fund through project IDIFEDER/2020/003.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the 6th Inter-*

national Conference on Learning Representations, Vancouver, Canada.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the 4th Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy.

Mauro Cettolo, Jan Niehues, Sebastian Stuker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *Proceedings of the 11th International Workshop on Spoken Language Translation*, pages 2–17, Lake Tahoe, USA.

Mauro Cettolo, Jan Niehues, Sebastian Stuker, Luisa Bentivogli, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, pages 2–14, Da Nang, Vietnam.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, USA.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 794–803, Stockholm, Sweden.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the 2nd Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544, Florence, Italy.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, USA.

Demi Guo, Yoon Kim, and Alexander Rush. 2020. Sequence-level mixed sample data augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5547–5552, Online.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada.

Katherine Lee, Orhan Firat, Ashish Agarwal, C. Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In *Interpretability and Robustness for Audio, Speech and Language, NIPS 2018 Workshop*, Montréal, Canada.

Guanlin Li, Lemao Liu, Guoping Huang, Conghui Zhu, and Tiejun Zhao. 2019. Understanding data augmentation in neural machine translation: Two perspectives towards generalization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5689–5695, Hong Kong, China.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Online.

Mohammad Norouzi, Samy Bengio, Navdeep Jaitly, Mike Schuster, Yonghui Wu, Dale Schuurmans, et al. 2016. Reward augmented maximum likelihood for neural structured prediction. *Advances In Neural Information Processing Systems*, 29:1723–1731.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3956–3965, Stockholm, Sweden.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618.

Matt Post. 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.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Accepted to NAACL 2021, Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Felipe Sánchez-Martínez, Rafael C. Carrasco, Miguel A. Martínez-Prieto, and Joaquín Adiego. 2012. Generalized biwords for bitext compression and translation spotting. *Journal of Artificial Intelligence Research*, 43:389–418.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the 1st Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Edinburgh neural machine translation systems for WMT16. In *Proceedings of the 1st Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *ArXiv preprint*, abs/2009.13818.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, USA.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions 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 1126–1140, Online.

Chaojun Wang and Rico Sennrich. 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.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Luxi Xing, and Weihua Luo. 2020. Uncertainty-aware semantic augmentation for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2724–2735, Online.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*, Toulon, France.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2019. Effectively training neural machine translation models with monolingual data. *Neurocomputing*, 333:240–247.

Huaao Zhang, Shigui Qiu, Xiangyu Duan, and Min Zhang. 2020. Token drop mechanism for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4298–4303, Barcelona, Spain (Online).

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

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Tong Xu. 2019. Regularizing neural machine translation by target-bidirectional agreement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 443–450, Honolulu, USA.

A Training details

The DA hyperparameters that maximized BLEU on the development set on the first training run are depicted in Table 6. Subsequent training runs were executed only with these best values. For the combination of SwitchOut and RAML, following Wang et al. (2018), firstly the best τ_x for SwitchOut was determined and, afterwards, the best τ_y for RAML was sought by fixing τ_x. For Switchout/RAML, the table depicts the value of τ^{-1}, while for the MTL DA tasks, it shows the value of α.

B Results with chrF++

Tables 7 and 8 show the chrF++ scores (Popović, 2017) for the in-domain automatic evaluations based on IWSLT parallel training data and on extended training data with back-translation, respectively; they are the counterpart to tables 3 and 4. Table 9 shows the chrF++ scores for the out-of-domain automatic evaluations whose corresponding BLEU scores are reported in Table 5. Scores were computed with `SacreBLEU` (Post, 2018).[15]

C Hallucinations

We motivate here the choice of thresholds used in the hallucination detection approach discussed in Sec. 5. Figure 3 shows an histogram with the normalized frequencies of the values of the adjusted BLEU with the concatenation of all the test data used (for English–German and German–English) in Fig. 2. It can be seen that more than 20% of the sentences would be regarded as hallucinations by our identification approach; our empirical observations corroborate this point, which may be explained by the low-resource scenario in which our experiments are run.

Table 10 shows some examples of references and generated translations together with the corresponding adjusted BLEU scores. It includes two output sentences which are regarded as hallucinations and one that is not.

The sentence-level smoothing approach used when computing the adjusted BLEU scores was based on the common technique (Chen and Cherry, 2014) of adding 1 to the matched n-gram count and the total n-gram count for n ranging from 2 to the maximum order of n-grams N (N is usually 4, but it is 2 in our case). Notice that this implies that if no unigram is matched, the resulting BLEU is 0. We thus consider that if no single token co-occurs, an hallucination is happening. However, bigram counts are smoothed as we do not want them to excessively affect the score. In fact, instead of adding 1 to both counts, we add 0.1. This is in line with weighting the precision ratio for bigrams with a weight (0.2) four times smaller than that of unigrams (0.8).

[15] Version string: `chrF2+lang.vi-en+numchars.6+space.false+version.1.5.0`

Hyperparam.	Task	en-de	de-en	en-he	he-en	en-vi	vi-en
τ_x^{-1}	SwitchOut	1.1	0.85	0.9	0.6	0.95	1.0
τ_y^{-1}	RAML	0.7	1.1	0.5	0.95	0.85	0.8
τ_y^{-1}	SwitchOut+RAML	0.7	1.0	0.95	1.0	0.5	0.8
α	swap	0.2	0.1	0.4	0.2	0.2	0.1
α	token	0.8	0.8	0.1	0.7	0.3	0.6
α	replace	0.2	0.5	0.3	0.2	0.3	0.2

Table 6: Data augmentation hyperparameters that maximized BLEU on the development set.

IWSLT parallel data only

Task	en-de	de-en	en-he	he-en	en-vi	vi-en
baseline	52.0 ± 0.1	53.1 ± 0.2	47.7 ± 0.2	54.6 ± 0.1	48.6 ± 0.3	49.2 ± 0.1
SwitchOut	52.2 ± 0.4	53.1 ± 0.2	47.6 ± 0.4	54.5 ± 0.3	48.1 ± 0.2	49.7 ± 0.6
RAML	52.5 ± 0.2	53.3 ± 0.2	48.2 ± 0.3	54.6 ± 0.1	48.7 ± 0.1	49.8 ± 0.3
SwitchOut+RAML	52.5 ± 0.1	53.3 ± 0.5	48.0 ± 0.1	54.6 ± 0.3	48.7 ± 0.3	49.6 ± 0.5
swap	52.4 ± 0.1	53.5 ± 0.1	48.1 ± 0.2	55.1 ± 0.2	48.5 ± 0.1	50.3 ± 0.3
token	52.6 ± 0.4	53.1 ± 0.4	47.7 ± 0.3	54.7 ± 0.4	48.9 ± 0.2	50.2 ± 0.1
source	52.4 ± 0.2	53.6 ± 0.4	47.5 ± 0.3	54.9 ± 0.3	49.2 ± 0.2	49.9 ± 0.4
reverse	53.1 ± 0.3	53.7 ± 0.2	48.4 ± 0.2	55.5 ± 0.1	49.1 ± 0.5	50.6 ± 0.0
mono	52.6 ± 0.3	53.8 ± 0.3	48.1 ± 0.1	54.9 ± 0.4	49.0 ± 0.1	49.9 ± 0.2
replace	$\mathbf{53.4 \pm 0.2}$	54.2 ± 0.2	48.6 ± 0.1	$\mathbf{55.8 \pm 0.4}$	49.5 ± 0.2	$\mathbf{50.7 \pm 0.5}$
reverse+replace	$\mathbf{53.7 \pm 0.4}$	$\mathbf{54.5 \pm 0.3}$	$\mathbf{49.3 \pm 0.2}$	$\mathbf{56.1 \pm 0.3}$	$\mathbf{49.9 \pm 0.4}$	$\mathbf{51.2 \pm 0.2}$
reverse+mono+replace	$\mathbf{53.8 \pm 0.3}$	$\mathbf{54.8 \pm 0.2}$	$\mathbf{49.3 \pm 0.2}$	$\mathbf{56.1 \pm 0.3}$	$\mathbf{50.0 \pm 0.2}$	$\mathbf{51.5 \pm 0.4}$

Table 7: Mean and standard deviation of the chrF++ scores obtained when translating in-domain test sets with the baseline system, three other reference systems, and our MTL DA approach, using different auxiliary tasks and combinations of them. Systems were trained only on IWSLT parallel data. The best results for each language pair, and those falling within one standard deviation from them, are highlighted in bold.

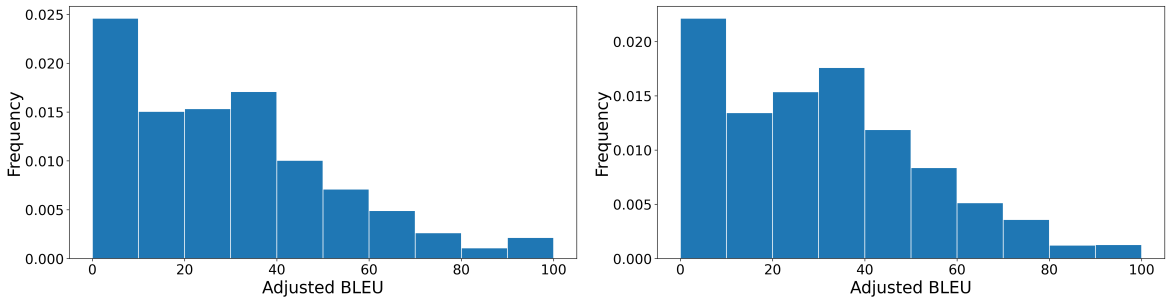
IWSLT parallel data + back-translated data

Task	de-en	he-en	vi-en
baseline	54.7 ± 0.4	56.6 ± 0.0	51.7 ± 0.2
SwitchOut+RAML	54.9 ± 0.6	56.3 ± 0.2	$\mathbf{52.2 \pm 0.1}$
reverse+mono+replace	$\mathbf{55.8 \pm 0.1}$	$\mathbf{57.4 \pm 0.2}$	$\mathbf{52.6 \pm 0.7}$

Table 8: Mean and standard deviation of the chrF++ scores obtained when translating in-domain test sets with the baseline system, the combination of SwitchOut and RAML, and the best combination of auxiliary task in our MTL DA approach. Systems were trained on a combination of parallel and back-translated data. The best results for each language pair, and those falling within one standard deviation from them, are highlighted in bold.

Domain	IT		Law		Medical	
Direction	en-de	de-en	en-de	de-en	en-de	de-en
baseline	25.4 ± 2.7	31.6 ± 4.4	33.4 ± 0.8	33.7 ± 0.4	34.6 ± 0.5	34.5 ± 0.8
SwitchOut	29.7 ± 3.2	29.7 ± 4.8	34.0 ± 0.4	32.8 ± 0.8	35.6 ± 0.7	34.3 ± 0.4
RAML	31.3 ± 1.8	31.8 ± 4.2	34.6 ± 0.5	34.2 ± 0.1	35.6 ± 0.6	36.0 ± 0.6
SwitchOut+RAML	33.3 ± 0.6	32.8 ± 3.1	34.1 ± 0.3	33.1 ± 0.6	36.3 ± 0.9	34.7 ± 0.4
reverse	35.1 ± 0.4	38.2 ± 0.2	34.6 ± 0.4	33.7 ± 0.6	36.9 ± 0.1	36.1 ± 0.3
reverse+mono+replace	$\mathbf{37.6 \pm 0.3}$	$\mathbf{39.7 \pm 0.7}$	$\mathbf{36.6 \pm 0.5}$	$\mathbf{35.5 \pm 0.8}$	$\mathbf{39.4 \pm 0.8}$	$\mathbf{38.5 \pm 0.4}$

Table 9: Mean and standard deviation of the chrF++ scores obtained when translating out-of-domain texts in the IT, law and medical domains. The best results for each language pair, and those falling within one standard deviation from them, are highlighted in bold.



(a) English–German (b) German–English

Figure 3: Adjusted BLEUs for the test data used in the hallucination analysis of Sec. 5.

Adjusted BLEU	Type	Sentence
1.74	input reference output	artikel 1 der verordnung (eg) nr. 1002/2004 erhält folgende fassung: article 1 of regulation (ec) no 1002/2004 shall be amended as follows: articles one of the figure-to-vis-it-vis-vis-a-vis-vis-vis-a-vis-vis-vis-vis-vis-vis-vis-vis-vis-vis-a-vis-vis-vis-vis-a-vis-vis-vis.
28.89	input reference output	artikel 1 der verordnung (eg) nr. 1002/2004 erhält folgende fassung: article 1 of regulation (ec) no 1002/2004 shall be amended as follows: articles 1 the requirement (eg) number 1002 / 2004 receives the following framework.
0	input reference output	verfalldatum expiry date dr: why don't you think about this?

Table 10: An output sentence, emitted by the baseline system, which was labelled as an hallucination (adjusted BLEU of 1.74). Below that, a non-hallucinated translation (adjusted BLEU of 28.89) generated by one of the DA systems. At the bottom, a more extreme example of hallucination with an adjusted BLEU of 0.