Joint Dropout: Improving Generalizability in Low-Resource Neural Machine Translation through Phrase Pair Variables

Ali Araabi a.araabi@uva.nl
Vlad Niculae v.niculae@uva.nl
Christof Monz c.monz@uva.nl
Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Abstract

Despite the tremendous success of Neural Machine Translation (NMT), its performance on low-resource language pairs still remains subpar, partly due to the limited ability to handle previously unseen inputs, i.e., generalization. In this paper, we propose a method called *Joint Dropout*, that addresses the challenge of low-resource neural machine translation by substituting phrases with variables, resulting in significant enhancement of compositionality, which is a key aspect of generalization. We observe a substantial improvement in translation quality for language pairs with minimal resources, as seen in BLEU and Direct Assessment scores. Furthermore, we conduct an error analysis, and find Joint Dropout to also enhance generalizability of low-resource NMT in terms of robustness and adaptability across different domains.

1 Introduction

Although Neural Machine Translation (NMT) has made remarkable advances (Vaswani et al., 2017), it still requires large amounts of data to induce correct generalizations that characterize human intelligence (Lake et al., 2017). However, such a vast amount of data to make robust, reliable, and fair predictions is not available for low-resource NMT (Koehn and Knowles, 2017).

The generalizability of NMT has been extensively studied in prior research, revealing the volatile behaviour of translation outputs when even a single token in the source sentence is modified (Belinkov and Bisk, 2018; Fadaee and Monz, 2020; Li et al., 2021). For instance, in the sentence "*smallpox killed billions of people on this planet*" from our IWSLT test set, when replacing the noun "*smallpox*" with another acute disease like "*tuberculosis*", the model should ideally generate a correct translation by only modifying the relevant part while keeping the rest of the sentence unchanged. However, in many instances, such a small perturbation adversely affects the translation of the entire sentence, highlighting the limited generalization and robustness of existing NMT models (Fadaee and Monz, 2020).

Compositionality is regarded as the most prominent form of generalization that embodies the ability of human intelligence to generalize to new data, tasks, and domains (Schmidhuber, 1990; Lake and Baroni, 2018), while other types mostly focus on the practical considerations across domains, tasks, and languages, model robustness, and structural generalization (Hupkes et al., 2022). Research in compositional generalization has two main aspects: evaluating the current models' compositional abilities as well as improving them.

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 12–25
September 4–8, 2023, Macau SAR, China.

In terms of evaluation, some studies use artificially created test sets that mimic arithmetic-like compositionality (Lake and Baroni, 2018), while others evaluate compositionality in a more natural way (Keysers et al., 2020; Kim and Linzen, 2020; Dankers et al., 2022). In terms of improvement, earlier work aimed to enhance the models' compositional abilities on tasks such as semantic parsing datasets (Qiu et al., 2022), math word problem solving (Lan et al., 2022), data-to-text generation (Mehta et al., 2022), and classification (Kim et al., 2021). As for NMT, previous work has shown shortcomings in systematic compositional skills Lake and Baroni (2018); Li et al. (2021), particularly for low-resource languages Dankers et al. (2022), yet no direct improvements have been proposed.

We aim to improve compositionality in NMT, with a focus on low-resource scenarios that necessitate more robustness to form new combinations of previously seen smaller units. To achieve this, we introduce Joint Dropout (JD), a simple and effective method that jointly replaces translation-equivalent phrase pairs in the source and target sentences with variables, encouraging the model to maintain the translation of the remaining sentence, regardless of the dropped phrases. JD is orthogonal to and compatible with other methods for improving NMT performance. Specifically, it is designed to be data-centric and model-agnostic, allowing it to be easily combined with existing techniques that focus on different aspects of the NMT pipeline.

Our analysis on simulated and real low-resource data demonstrates JD's ability to significantly improve compositional generalization and translation quality.

2 Methodology

Generalization has been a longstanding concern in the field of machine translation. In the past, Statistical MT utilized phrases as the fundamental translation units in order to consider contextual information, such as in Phrase-Based Statistical Machine Translation (Zens et al., 2002, PBSMT). To increase generalization, Hierarchical PBSMT proposed by Chiang (2005) builds upon the bilingual phrase pairs of PBSMT to learn hierarchical rules, capturing discontinuous translation equivalences and therefore allowing for better generalization.

Similarly, JD leverages bilingual phrases to make the rest of the translation not dependent on a specific phrase pair. However, the main idea behind JD originates from compositionality: the meaning of a sentence is a function of the meanings of its known atoms and how they are systematically and syntactically combined (Partee et al., 1984). By substituting *meaning* with *translation* in this definition, we come up with a rule of compositionality for translation systems:

$$\tau(P \circ Q) = \tau(P) \circ \tau(Q) \tag{1}$$

in which τ is the translation function, P and Q are the constituents of the sentence, and $\circ$ is a combiner. JD aims to transfer the principle of compositionality to the translation model in order to improve generalization and robustness of NMT by replacing joint phrases with variables. To exemplify, given the De-En sentence pair $\langle$*Sie hat Rom besucht, She visited Rome*$\rangle$, we replace nouns with variables: $\langle X_1$ *hat* X_2 *besucht,* Y_1 *visited* $Y_2\rangle$. Per Equation1:

$$\begin{aligned}
&\tau(\text{Sie hat Rom besucht}) \\
&= \tau(((X_1 \text{ hat } X_2 \text{ besucht}) \circ_{X_1} \text{ Sie}) \circ_{X_2} \text{Rom}) \\
&= \tau((X_1 \text{ hat } X_2 \text{ besucht}) \circ_{X_1} \text{Sie}) \circ_{\tau(X_2)} \tau(\text{Rom}) \\
&= (\tau(X_1 \text{ hat } X_2 \text{ besucht}) \circ_{\tau(X_1)} \tau(\text{Sie})) \circ_{\tau(X_2)} \tau(\text{Rom}) \\
&= ((Y_1 \text{ visited } Y_2) \circ_{Y_1} \text{ She}) \circ_{Y_2} \text{Rome} \\
&= \text{She visited Rome}
\end{aligned} \tag{2}$$

where $\tau(X_i) = Y_i$, and $\sigma \circ_X \gamma = \sigma[X_i \backslash \gamma]$, i.e., $\circ_X$ performs the replacement of γ in the position X_i in the sentence σ. In the above sketch, we disregard any potential dependencies

within the sentence. However, the variables are independent of the rest of the sentence in any manner. Therefore, our goal is to enable the model to translate the entire sentence without being affected by the specific words or phrases at position X_i. Hence, if the model learns the rules of composition properly, changing one or more lexical units will not hurt the rest of the translation. To this end, inspired by hierarchical PBSMT, we make use of bilingual phrases to improve generalization in low-resource NMT. However, since NMT has a strong capability to learn ordering through the cross-attention mechanism (Toral and Sánchez-Cartagena, 2017), our aim is not to directly apply hierarchical PBSMT to NMT, but to propose an approximation as a lightweight and efficient regularization method.

First, using Eflomal (Östling and Tiedemann, 2016), an efficient word alignment tool, we generate symmetrized word alignments for the parallel training corpus to find the correspondences between source and target words in each pair of training sentences. Then, we use alignments as the input to generate the phrase translation table by decomposing the source and target sentences into a set of dozens of bilingual phrase pairs that are consistent with the word alignment (Koehn et al., 2003). In the next step, we select phrase pairs from the phrase table for each pair of training sentences and replace them with joint variables of (X_i,Y_i). More specifically, given a pair of sentences $S = \{w_1, w_2, ...w_n\}$ and $T = \{w'_1, w'_2, ...w'_m\}$, after substitution the sentences are $S = \{w_1, ..., X_i, ..., w_l, ..., X_j, ..., w_n\}$ and $T = \{w'_1, ..., Y_i, ..., w'_k, ..., Y_j, ..., w'_m\}$, where X and Y are variables corresponding to the source and target phrases, respectively. We discuss different criteria to replace phrases with variables in §3.2.[1] Finally, we add the variable-induced corpus to the original training set, effectively doubling its size.[2]

3 Experiments

In this section, we present a comprehensive overview of our experiments. We begin by providing details regarding the datasets used and the training systems employed. Next, we delve into the specific criteria we considered when replacing phrases with variables. Subsequently, we discuss the significant improvements achieved by our proposed method, JD, across various aspects, including compositional generalization, translation performance, robustness, and the ability to generalize across domains.

3.1 Experimental setup

Data. For the preliminary experiments, we use the TED data from the IWSLT 2014 German-English (De-En) shared translation task (Cettolo et al., 2014) and randomly sample from the training data to represent various low-resource settings. In order to evaluate the models trained on IWSLT subsets, we use the concatenation of the IWLST 2014 dev sets (tst2010–2012, dev2010, dev2012) as our test set, which consists of 6,750 sentence pairs.

We further evaluate JD on multiple actual low-resource language pairs: Belarusian (Be), Galician (Gl), and Slovak (Sk) TED talks (Qi et al., 2018) and Slovenian (Sl) from IWSLT2014 (Cettolo et al., 2014) with training sets ranging from 4.5k to 55k sentence pairs.

In order to evaluate the compositional ability of JD, following Dankers et al. (2022), we use an English-Dutch (En-Nl) training set from OPUS [3] (Tiedemann and Thottingal, 2020) and randomly sample to create low-resource sets. To evaluate these models, we use both the 'dev' and the 'devtest' sets from FLORES-101 (Goyal et al., 2022) as the validation and test data.

[1]The code is available at https://github.com/aliaraabi/Joint_Dropout

[2]We ensure all models undergo the same maximum number of updates during training, allowing a fair evaluation.

[3]Available on https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/master/data/README-v2020-07-28.md

Setup	#Phrases	BLEU
T-base	0	12.2
T-opt.	0	18.0
T-opt. + JD_PP	8013	18.6
T-opt. + JD_VP	8013	18.8
T-opt. + JD_NP	8013	18.7
T-opt. + JD_Mix	8013	18.8

Table 1: Results of Transformer-base, Transformer-optimized and Joint Dropout with various phrase types on 10K De-En training samples. Noun Phrases (NP), Prepositional Phrases (PP), Verb Phrases (VP), and mixture (Mix) of all the above.

Setup	BLEU
T-opt.	18.0
T-opt. + JD	19.9
T-opt. + target variables only	15.5
T-opt. + source variables only	17.3
T-opt. + not aligned variables	17.8

Table 2: Importance of jointly dropping aligned phrases for model trained on 10K De-En samples.

Pre-processing. We apply punctuation normalization, tokenization, data cleaning, and truecasing using the Moses scripts (Koehn et al., 2007). The sentence length is limited to a maximum of 175 tokens during training. After replacing phrases with variables, we also apply BPE segmentation (Sennrich et al., 2016b) with the parameter tailored to the low-resource training data (Araabi and Monz, 2020). We ensure that variables are not split into smaller segments.

Data annotation. To generate a realistic test set for evaluating robustness against sentence perturbation, we first randomly select 300 translation outputs from the inference stage of baseline systems trained using optimized parameters on 20k samples. These outputs are then ranked using the Direct Assessment (DA) method by engaging native annotators. The top 100 outputs are then selected and the corresponding outputs from the model trained with JD are extracted and ranked using DA. Next, the input sentences are modified by replacing specific phrases or words while ensuring their syntactic and semantic accuracy. After obtaining the outputs for both the baseline and JD systems on the perturbed sentences, we conduct a DA on them.

Training system. To conduct our experiments, we employ two different models: Transformer-optimized (Araabi and Monz, 2020), specifically tailored to low-resource NMT and Transformer-base with its default hyper-parameters (Vaswani et al., 2017). This choice allows us to demonstrate that the improvements achieved are consistent and independent of the specific model settings. We use the Fairseq library (Ott et al., 2019) for our experiments and average sacreBLEU[4] (Post, 2018) over three runs as the evaluation metric. All of the models are trained on a single GPU for a few hours with the model parameters ranging from 28M to 47M.

3.2 Joint Dropout parameters

The following conditions are considered in replacing phrases with variables. First, we do not allow two adjacent phrases to be replaced with variables. Although phrases can vary in length, we consider all phrases as potential candidates for substitution with variables, irrespective of their length. After conducting initial experiments, we have determined that setting the maximum number of variables allowed in each sentence to 10 yields satisfactory results.

Since noun phrases are the most cross-linguistically common phrases, we hypothesize that they are good candidates to be replaced. Therefore, in a set of experiments we investigate the choice of phrase types. We consider four different scenarios: replacing 1) only Noun

[4]sacreBLEU parameters: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

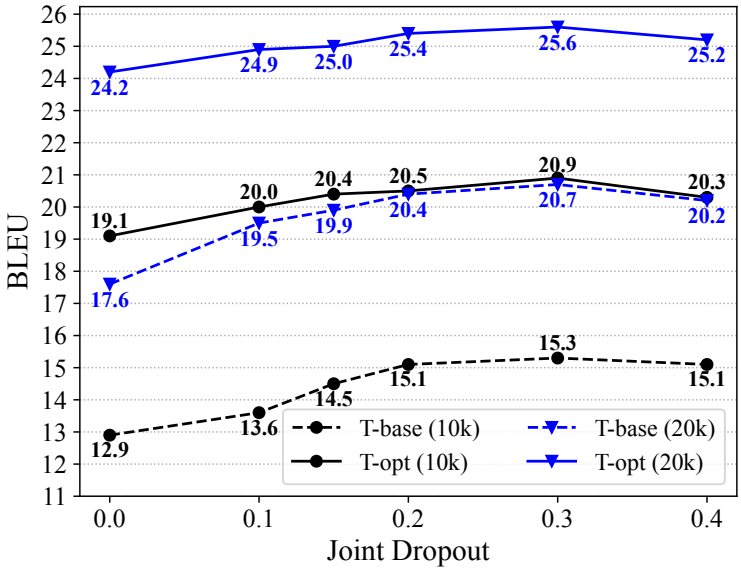

Figure 1: Effect of different Joint Dropout rates on Transformer-base and Transformer-optimized, on the validation sets of two De-En training subsets.

Phrases (NP), 2) only Prepositional Phrases (PP), 3) only Verb Phrases (VP), and 4) mixtures of all the above. We train four systems on 10k samples from the TED talks dataset with four different substitution scenarios yet the same number of variables (8013).[5] We use the constituency parser from Stanford CoreNLP (Manning et al., 2014).

It is important to note that our selection of phrase pairs in both languages is solely based on English constituency parse trees. We do not rely on the use of a constituency parser, which is often not available for many low-resource languages. The results presented in Table 1 demonstrate that the choice of different phrase types does not lead to significant differences in our method. Therefore, our approach eliminates the need for a constituency parser, making it applicable to a wider range of low-resource languages. For the rest of the experiments, we substitute phrases regardless of their types.

To make JD independent of a phrase translation table, we consider not-aligned phrases in both or either translation sides. The importance of using aligned phrases is demonstrated in Table 2, where it is observed that utilizing not-aligned phrases results in a degradation of performance by 2.5 BLEU points. This finding highlights the significance of incorporating aligned phrases in the JD method.

To maintain control over the number of variables across the entire training corpus, we introduce a concept called the *Joint Dropout rate*. This rate is determined by calculating the proportion of dropped tokens, specifically from within phrases, in relation to the total length of both the source and target sentences. By utilizing this Joint Dropout rate, we can effectively regulate and manage the presence of variables throughout the training process. Figure 1 illustrates the improvements achieved by two distinct systems as the Joint Dropout rate increases. Notably, JD consistently improves the performance of both the Transformer-base and Transformer optimized models. Specifically, on a dataset of 110k samples, JD yields a notable increase of +2.4 BLEU points for the Transformer-base model and +1.8 BLEU points for the

[5] 8013 is the number of all possible substitutions for PPs.

	BLEU		Consistency	
#Samples	T-opt.	T.opt.+JD	T-opt.	T.opt.+JD
$5k$	4.2	**6.1**	2.0	**4.0***
$20k$	10.4	**10.7**	8.1	**11.0***
$40k$	12.8	**13.4**	13.1	**15.6***
$80k$	**16.4**	**16.4**	37.1	**43.8***
$200k$	**19.2**	18.7	58.2	**65.4***

Table 3: BLEU and consistency scores (En $\rightarrow$ Nl) when replacing a noun in the subject position with a different noun. Significant improvements on compositionality of JD over the strong baseline are marked with $*$ (approximate randomization, $p < 0.01$).

Transformer-optimized model. Moreover, when evaluating a larger dataset of 20k samples, JD further improves translation quality by $+3.1$ BLEU points for the Transformer-base model and $+1.4$ BLEU points for the Transformer-optimized model.

We see that the Joint Dropout rate of 0.3 is a good choice, while more noise in the training set hurts performance. We use this rate for the remainder of the experiments.

3.3 Compositional generalization

Unlike phenomena such as idioms, which require a more global understanding, JD concentrates on improving compositionality at the local level. In this section, we aim to evaluate our method on local compositionality. Here, we take advantage of the most relevant theoretically grounded test from Hupkes et al. (2020) which is *systematicity*, a notion frequently used in the context of compositionality. This attribute of the model concerns the recombination of known parts and rules, ensuring that the model's ability to grasp novel inputs is systematically tied to their aptitude to comprehend related inputs. For instance, understanding "smallpox killed billions of people on this planet" and "tuberculosis", also implies understanding "tuberculosis killed billions of people on this planet".

Given that there are an infinite number of potential novel combinations that can be derived from known parts in natural data, we concentrate on a sentence-level, context-free rule: S $\rightarrow$ NP VP, as proposed by Dankers et al. (2022), where a noun from the NP in the subject position is replaced with a different noun, while maintaining number agreement with the VP. Additionally, they highlight that a systematic system necessitates consistency. We assess this systematicity of translations based on their consistency across various contexts when presenting words or phrases. Consistency is measured by evaluating the equality between two translations while taking into account anticipated modifications. In S $\rightarrow$ NP VP setup, after replacement, translations are deemed consistent if there is only one word difference between them. Table 3 illustrates that JD consistently enhances the consistency scores for various low-resource data conditions.

3.4 Translation performance

In this section, we conduct a comprehensive evaluation of translation quality across multiple language pairs to assess the effectiveness of JD. The results presented in Table 4 highlight the significant improvements in translation quality achieved by JD for actual low-resource language pairs. Importantly, these improvements also hold true for the reverse language direction.

Furthermore, we compare JD to three comparable methods for dropping tokens: Zero-Out, where the embedding of a token is set to zero (Sennrich et al., 2016a), Token Drop, which replaces tokens with the `<dropped>` tag Zhang et al. (2020), and SwitchOut, where words are replaced with random words from their corresponding vocabularies Wang et al. (2018). The

Method	Be-En	Gl-En	Sl-En	Sk-En	En-Be	En-Gl	En-Sl	En-Sk
T-base	4.6	13.4	8.9	24.0	3.5	10.1	6.8	19.0
T-base + JD	**6.5**	**15.8**	**10.2**	**25.0**	**4.5**	**12.9**	**7.8**	**19.2**
T-opt.	8.0	21.8	15.2	28.9	5.5	18.3	12.3	23.1
T-opt. + JD	**9.9**	**22.8**	**16.1**	**29.8**	**7.3**	**18.9**	**12.7**	**23.5**

Table 4: BLEU scores for actual extremely low-resource languages: Be, Gl, Sl, and Sk with 4.5k, 10k, 13k, and 55k training samples, respectively.

Method	5k	10k	20k
T-opt.	13.4	18.0	23.0
T-opt. + ZO	13.6	18.3	22.8
T-opt. + TD	9.5	16.8	23.9
T-opt. + SW	13.4	18.4	24.0
T-opt. + JD	**15.2**	**19.9**	**24.4**

(a) Transformer-optimized

Method	5k	10k	20k
T-base	8.6	12.1	16.6
T-base + ZO	8.9	13.3	18.3
T-base + TD	5.3	8.9	14.6
T-base + SW	5.5	9.8	14.5
T-base + JD	**9.8**	**14.5**	**19.1**

(b) Transformer-base

Table 5: Comparing BLEU scores for Joint Dropout (JD) and the reimplementations of Token Drop (TD), Zero Out (ZO), and SwitchOut (SW) on 5k, 10k and 20k training samples from IWSLT De-En.

results in Table 5a demonstrate that Zero-Out only provides marginal improvements. Moreover, both Token Drop and SwitchOut methods prove to be ineffective in low-resource scenarios. In contrast, JD consistently outperforms these methods, particularly in extreme low-resource cases. As shown in Table 5a, Zero-Out only provides marginal improvements. In addition, while Token Drop and SwitchOut methods prove to be ineffective in low-resource situations, JD consistently yields the largest improvements, especially for extreme low-resource cases. In addition, Table 5b provides additional evidence supporting the superiority of JD over similar methods, even when optimized parameters for the Transformer model are not specifically chosen.

3.5 NMT Robustness

Recent work has shown that trivial modifications to the source sentence can cause unexpected changes in the translation (Fadaee and Monz, 2020). Furthermore, models with stronger compositional abilities are anticipated to generate more robust translations Dankers et al. (2022). To evaluate the robustness of JD against such modifications, we differ from previous methods that automatically introduce noise to the test set (Michel and Neubig, 2018; Cheng et al., 2019) which is prone to creating semantic and syntactic errors in the input. Instead, we manually develop a more realistic test set.

First, based on Direct Assessment (DA) on a 100-point scale (Graham et al., 2013), we select the top 100 sentences out of randomly selected 300 translation outputs generated by a Transformer-optimized model trained on 20k samples. We then alter the input sentences by replacing a specific phrase or word, while ensuring that they remain syntactically and semantically accurate. Table 6 illustrates that perturbing the original sentences results in a smaller performance decrease for the model trained with JD, when compared to the baseline. This means that our proposed method significantly decreases the volatile behavior of low-resource NMT.

Table 7 shows an example of perturbing a sentence. After replacing "*ein Kind in Indien*" in

Method	Metric	Orig.	Per.	Δ
T-base	DA	62.1	49.3	−12.8
	BLEU	28.5	26.0	−2.5
T-base + JD	DA	69.8	59.3	**−10.5**
	BLEU	30.7	30.4	**−0.3**
T-opt.	DA	79.9	56.6	−23.3
	BLEU	37.4	31.8	−5.6
T-opt. + JD	DA	83.7	77.4	**−6.3**
	BLEU	41.8	39.9	**−1.9**

Table 6: Direct assessment and BLEU scores, pre and post input perturbation on random samples from De-En test set.

	Original test sentence	Test sentence after perturbation
Src	[**ein Kind in Indien**] sagt: "heute habe ich einen Affen gesehen".	{**meine Oma in China**} sagt: "heute habe ich einen Affen gesehen".
Ref.	[a child in India] says , "I saw a monkey today ."	{my grandmother in China} says, "I saw a monkey today ."
T-opt.	[a child in India] says, "today I've seen a monkeys."	{my grandmother's mother in China} says, "Look today."
T-opt. + JD	[a kid in India] says, "I've seen a monkeys today."	{my grandmother in China} says, "today I've seen a monkeys."

Table 7: By replacing the German noun phrase ***ein Kind in Indien*** [a child in India] with ***meine Oma in China*** [my grandmother in China], there is no undesirable behavior in the rest of the translation when using Joint Dropout. Underlined text means the rest of the translation is approximately the same with the reference, while the wavy underline means it has changed. Bracket shows the phrase that we perturb, while the curly bracket is the perturbed phrase

the source sentence with "*meine Oma in China*", while the rest of the translation is negatively affected using the baseline model, the JD shows more robustness against the input perturbation and does not exhibit any negative behavior.

3.6 Generalization across domains

In low-resource language settings, NMT systems frequently encounter challenges when it comes to achieving effective translation across distinct domains. This is primarily attributed to their tendency to prioritize the idiosyncrasies of the training domain, rather than capturing the broader linguistic characteristics shared by the language pairs. Therefore, in addition to evaluating generalization in terms of compositionality and robustness, it is also crucial to assess generalization concerning distributional shift and uncertainty estimation (Hupkes et al., 2022). While the definition of a domain is not precisely defined (van der Wees et al., 2015),for our evaluation, we consider TED talks and news as belonging to different domains.

Table 8 provides insights into the behavior of JD when there is a domain shift between the training domain (TED talks) and the test domain (news from WMT). The results demonstrate that JD exhibits greater robustness in such scenarios, showcasing its ability to better handle

Method	10k	20k	40k
T-base	2.4	3.9	7.1
T-base + JD	**3.2**	**5.4**	**9.8**
T-opt.	6.2	8.7	13.9
T-opt. + JD	**7.5**	**10.9**	**14.6**

Table 8: Results of training on different subsamples of TED talks and testing on a domain with different distribution (Newstest2020).

distributional shifts and improve translation quality across different domains. This highlights the effectiveness of JD in mitigating the negative effects of domain-specific training and enhancing the generalizability of NMT systems in low-resource language pairs.

4 Conclusion

Despite the fact that NMT's success is closely tied to having large amounts of training data, it is still beneficial to explore methods that can help improve generalization when working with limited data. In this paper, we introduce Joint Dropout as a straightforward yet effective approach to enhancing the compositional generalization and translation quality of low-resource NMT. Specifically, we demonstrate that jointly replacing phrases with variables has a regularizing effect that mitigates overfitting by enabling the system to translate sentences regardless of the specific phrases present at the variable positions.

5 Future work

We only focus on improving generalizability of low-resource NMT, while higher-resource settings might also gain from joint variables. Additionally, we demonstrate the effectiveness of our proposed method using multiple low-resource language pairs, whereas there are many other language pairs with limited data. Furthermore, since JD tries to capture the rules of compositionality in translation, we expect more benefit to the language pairs with less similarity. Additionally, our approach is data-centric and model-agnostic, applicable to various models and tasks beyond the methods evaluated in this paper. Therefore, it has the potential to improve existing pre-trained models such as mBART (Liu et al., 2020), when fine-tuning on low-resource languages, but further experimentation is needed to confirm its effectiveness. We leave these investigations to future work.

6 Broader Impact

The implementation of NMT has brought about significant progress in the translation field, however, it also poses potential challenges such as liability for mistakes made by using NMT and mistranslation, which could be more of a concern when dealing with limited data. Furthermore, the high ability of NMT to generalize well presents a potential risk of difficulty in identifying errors, specifically those related to compositionality. This can be a concern in safety-critical domains where a single error can have severe consequences. Moreover, the ability of NMT to produce more coherent and fluent translations may impede the identification of where the system is malfunctioning, thus hindering the correction of errors or biases in the model.

References

Araabi, A. and Monz, C. (2020). Optimizing transformer for low-resource neural machine translation. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3429–3435. International Committee on Computational Linguistics.

Belinkov, Y. and Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2014). Report on the 11th IWSLT evaluation campaign. In Federico, M., Stüker, S., and Yvon, F., editors, *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign@IWSLT 2014, Lake Tahoe, CA, USA, December 4-5, 2014*.

Cheng, Y., Jiang, L., and Macherey, W. (2019). Robust neural machine translation with doubly adversarial inputs. In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4324–4333. Association for Computational Linguistics.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In Knight, K., Ng, H. T., and Oflazer, K., editors, *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 263–270. The Association for Computer Linguistics.

Dankers, V., Bruni, E., and Hupkes, D. (2022). The paradox of the compositionality of natural language: A neural machine translation case study. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4154–4175. Association for Computational Linguistics.

Fadaee, M. and Monz, C. (2020). The unreasonable volatility of neural machine translation models. In Birch, A., Finch, A. M., Hayashi, H., Heafield, K., Junczys-Dowmunt, M., Konstas, I., Li, X., Neubig, G., and Oda, Y., editors, *Proceedings of the Fourth Workshop on Neural Generation and Translation, NGT@ACL 2020, Online, July 5-10, 2020*, pages 88–96. Association for Computational Linguistics.

Goyal, N., Gao, C., Chaudhary, V., Chen, P., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguistics*, 10:522–538.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In Dipper, S., Liakata, M., and Pareja-Lora, A., editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, LAW-ID@ACL 2013, August 8-9, 2013, Sofia, Bulgaria*, pages 33–41. The Association for Computer Linguistics.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. (2020). Compositionality decomposed: How do neural networks generalise? *J. Artif. Intell. Res.*, 67:757–795.

Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., Ulmer, D., Schottmann, F., Batsuren, K., Sun, K., Sinha, K., Khalatbari, L., Ryskina, M., Frieske, R., Cotterell, R., and Jin, Z. (2022). State-of-the-art generalisation research in NLP: a taxonomy and review. *CoRR*, abs/2210.03050.

Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, L., Tihon, T., Tsarkov, D., Wang, X., van Zee, M., and Bousquet, O. (2020). Measuring compositional generalization: A comprehensive method on realistic data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kim, J., Ravikumar, P., Ainslie, J., and Ontañón, S. (2021). Improving compositional generalization in classification tasks via structure annotations. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 637–645. Association for Computational Linguistics.

Kim, N. and Linzen, T. (2020). COGS: A compositional generalization challenge based on semantic interpretation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9087–9105. Association for Computational Linguistics.

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 Carroll, J. A., van den Bosch, A., and Zaenen, A., editors, *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In Luong, T., Birch, A., Neubig, G., and Finch, A. M., editors, *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In Hearst, M. A. and Ostendorf, M., editors, *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics.

Lake, B. M. and Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888. PMLR.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Lan, Y., Wang, L., Jiang, J., and Lim, E. (2022). Improving compositional generalization in math word problem solving. *CoRR*, abs/2209.01352.

Li, Y., Yin, Y., Chen, Y., and Zhang, Y. (2021). On compositional generalization of neural machine translation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4767–4780. Association for Computational Linguistics.

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

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mehta, S. V., Rao, J., Tay, Y., Kale, M., Parikh, A., and Strubell, E. (2022). Improving compositional generalization with self-training for data-to-text generation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4205–4219. Association for Computational Linguistics.

Michel, P. and Neubig, G. (2018). MTNT: A testbed for machine translation of noisy text. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 543–553. Association for Computational Linguistics.

Östling, R. and Tiedemann, J. (2016). Efficient word alignment with markov chain monte carlo. *Prague Bull. Math. Linguistics*, 106:125–146.

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 Ammar, W., Louis, A., and Mostafazadeh, N., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Partee, B. et al. (1984). Compositionality. *Varieties of formal semantics*, 3:281–311.

Post, M. (2018). A call for clarity in reporting BLEU scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M. L., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? In Walker, M. A., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 529–535. Association for Computational Linguistics.

Qiu, L., Shaw, P., Pasupat, P., Nowak, P. K., Linzen, T., Sha, F., and Toutanova, K. (2022). Improving compositional generalization with latent structure and data augmentation. In Carpuat, M., de Marneffe, M., and Ruíz, I. V. M., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4341–4362. Association for Computational Linguistics.

Schmidhuber, J. (1990). Towards compositional learning in dynamic networks.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 371–376. The Association for Computer Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT - building open translation services for the world. In Forcada, M. L., Martins, A., Moniz, H., Turchi, M., Bisazza, A., Moorkens, J., Arenas, A. G., Nurminen, M., Marg, L., Fumega, S., Martins, B., Batista, F., Coheur, L., Escartín, C. P., and Trancoso, I., editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 479–480. European Association for Machine Translation.

Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 1063–1073. Association for Computational Linguistics.

van der Wees, M., Bisazza, A., Weerkamp, W., and Monz, C. (2015). What's in a domain? analyzing genre and topic differences in statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 560–566. The Association for Computer Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wang, X., Pham, H., Dai, Z., and Neubig, G. (2018). Switchout: an efficient data augmentation algorithm for neural machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 856–861. Association for Computational Linguistics.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In Jarke, M., Koehler, J., and Lakemeyer, G., editors, *KI 2002: Advances in Artificial Intelligence,*

25th Annual German Conference on AI, KI 2002, Aachen, Germany, September 16-20, 2002, Proceedings, volume 2479 of *Lecture Notes in Computer Science*, pages 18–32. Springer.

Zhang, H., Qiu, S., Duan, X., and Zhang, M. (2020). Token drop mechanism for neural machine translation. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4298–4303. International Committee on Computational Linguistics.