Deterministic Reversible Data Augmentation for Neural Machine Translation

Anonymous ACL submission

Abstract

Data augmentation is an effective way to diversify corpora in machine translation, but previous methods may introduce semantic inconsistency between original and augmented data because of irreversible operations and random subword sampling procedures. To generate both symbolically diverse and semantically consistent augmentation data, we propose Deterministic Reversible Data Augmentation (DRDA), a simple but effective data augmentation method for neural machine translation. DRDA adopts deterministic segmentations and reversible operations to generate multi-granularity subword representations and pulls them closer together with multi-view techniques. With no extra corpora or model changes required, DRDA outperforms strong baselines on several translation tasks with a clear margin (up to 4.3 BLEU gain over Transformer) and exhibits good robustness in noisy, low-resource, and cross-domain datasets. [1]

1 Introduction

Recent neural machine translation (NMT) models have led to dramatic improvements in translation quality. However, the powerful learning and memorizing ability of these models also leads to poor generalization and vulnerability to small perturbations like misspelling and paraphrasing (Belinkov and Bisk, 2017; Cheng et al., 2020).

A common solution to perturbation vulnerability is data augmentation (Sennrich et al., 2016b; Cheng et al., 2016), which is to create massive virtual training data with diverse symbolic representations under the premise of ensuring semantic consistency (Cheng et al., 2019, 2020). Symbolic diversity emphasizes that original and augmented data should differ significantly in token sequences, and semantic consistency requires that the two should be semantically similar. Previous data augmentation methods employ irreversible substitutions, like direct dropping or replacing discrete tokens to generate diverse data (Figure 1 A). Despite being able to improve data diversity, these augmentation operations are not reversible, and will inevitably introduce semantic loss to original texts, thus compromising the semantic consistency between original and augmented data.

Yet another way to generate diverse augmentation data without employing irreversible operations is subword regularization (Kudo, 2018; Provilkov et al., 2020). Subword regularization adopts random segmentations to sample subwords probabilistically thus generating diverse data. These methods are reversible because of the inherent reversibility of segmentations. However, due to the random sampling procedure of segmentation, they may adopt inappropriate subword segmentations (e.g., "`sup erm ark et`" in Figure 1 B). These sub-optimal segmentations may result in semantic perturbations and do damage to semantic consistency.

To summarize, previous methods have difficulty in completely retaining the semantics from corruption when diversifying the texts because of irreversible augmentation operations and probabilistic subword sampling.

To generate symbolically diverse and semantically consistent data, we propose Deterministic Reversible Data Augmentation (DRDA), a simple but effective augmentation approach. DRDA augments source sentences with their token representations in different granularities as shown in Figure 1 C. These representations are symbolically diverse, but also syntactically correct and semantically complete thanks to the reversible and deterministic segmentations in the multi-granularity segmentation process. To make full use of the semantic identity among all multi-granularity representations of one sentence, we also leverage the multi-view techniques in training to pull these representations closer together.

[1] The code will be released at `Anonymous Link`.

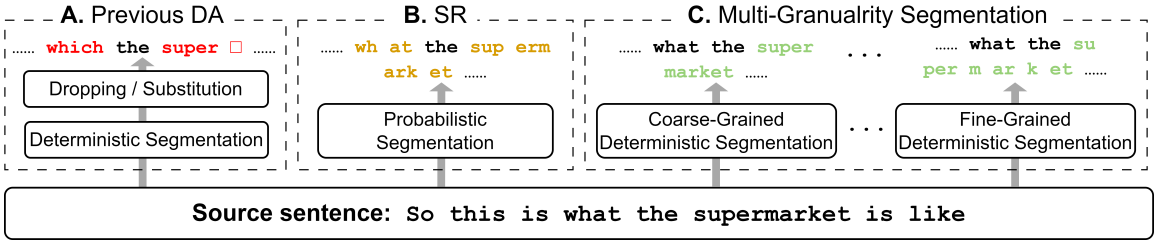

Figure 1: Subword piece sequences generated by previous data augmentation (A), subword regularization (B), and multi-granularity segmentation (C) representing the same source sentence. □ denotes an empty subword (a zero vector). Previous data augmentation methods result in semantic loss (red texts), subword regularization may sample inappropriate subwords (yellow texts), while multi-granularity segmentation generates symbolically diverse and semantically consistent augmentation data (green texts).

We conduct extensive experiments of different languages and scales and find that DRDA gains consistent improvements over strong baselines with clear margins. To further understand the factors that make DRDA work, we conduct insightful analyses of the effects DRDA imposed on semantic consistency, subword frequency, and subword semantic composition. We combine the empirical and theoretical verification of the consistency and offer a subword-level explanation of the mechanism of multi-granularity segmentations and multi-view techniques.

Our contributions are summarized as follows:

- We propose DRDA that exclusively employs deterministic reversible operations to generate diverse augmentation data without introducing semantic noise.
- We conduct extensive experiments and verify the high effectiveness of DRDA.
- To investigate the factors that make DRDA work, we combine empirical and theoretical analyses and offer insightful explanations.

2 Related Work

Augmentation methods Data augmentation can be categorized into back-translation like methods (Sennrich et al., 2016b; Edunov et al., 2018; Nguyen et al., 2020) and token substitution methods. DRDA is an instance of the latter category.

Several substitution methods uniformly select a word or token in a sentence and perform deletion or substitution (Zhang et al., 2020; Shen et al., 2020; Wang et al., 2018b; Norouzi et al., 2016; Gao et al., 2022). Cheng et al. (2019, 2020) constrained the substitution of a word in a small subset of synonyms, thus improving the semantics consistency. Kambhatla et al. (2022b) viewed the original corpus as plain text and applies a rotation encryption as data augmentation. Unlike previous methods, introducing multi-granularity takes advantage of the reversible nature of segmentation and causes no semantic loss.

Subword regularization The de-facto subword method, BPE (Sennrich et al., 2016c), still suffers from sub-optimality (Bostrom and Durrett, 2020). To overcome this sub-optimality, several subword regularization approaches are proposed. Kudo (2018) and Provilkov et al. (2020) presented subword regularization by modelling segmentation ambiguity. Wang et al. (2021) integrated BPE and BPE-Drop by enforcing the consistency using multi-view subword regularization, Wu et al. (2020) and Kambhatla et al. (2022a) combined BPE in `SentencePiece` and `subword-nmt` together to obtain regularization effects. DRDA is distinct from all the random sampling segmentation methods, as the augmentation data is generated deterministically. The determinism helps alleviate less reasonable segmentation, while achieving regularization effects as well.

In addition, other researches put efforts into taking advantage of multi-granularity representations, which can also be viewed as a subword regularization. Li et al. (2020) and Gao et al. (2020) adopted word lattice and convolutions of different kernel sizes respectively, Chen et al. (2018) and Li et al. (2022) combined levels of representation scales, Hao et al. (2019) modified self-attention module to introduce phrase modeling. Unlike these methods, DRDA requires no modification to model architectures and can be applied to universal tasks.

3 Background: Subword Segmentation

Subword segmentation models the probability of token sequence $\mathbf{x} = x_1, x_2, ..., x_m$ given a source

sentence $\mathbf{s}$. Previous deterministic subword segmentations choose the most probable sample:

$$\begin{aligned}\mathbf{x}^* &= \arg\max_{\mathbf{x}} P_{seg}(\mathbf{x}|\mathbf{s};p) \\ &= \arg\max_{\mathbf{x}\in V_p} P_{seg}(\mathbf{x}|\mathbf{s}),\end{aligned} \tag{1}$$

where p is the size of the vocabulary (a set of subword candidates), and each token x_i ($i \in \{1,2,...,m\}$) is selected from vocabulary V_p. For example, Byte Pair Encoding (BPE) assigns $P(\hat{\mathbf{x}}|\mathbf{s};p) = 1$ when $\hat{x}$ is obtained from the greedy merge process (Sennrich et al., 2016c).

To generate different segmentations for one word, subword regularization methods draw a segmentation from the segmentation distribution probabilistically:

$$\mathbf{x} \sim P_{seg}(\mathbf{x}|\mathbf{s};p). \tag{2}$$

For example, Kudo (2018) makes use of a unigram language model to sample segmentations on, and Provilkov et al. (2020) randomly interrupts the BPE merging process to generate multiple segmentations.

4 Deterministic Reversible Data Augmentation

Previous data augmentation and subword regularization approaches take irreversible operation (like discrete token substitution) and probabilistic segmentation sampling, which may introduce semantic loss or inappropriate subwords, thus affecting the semantic consistency. Our objective is to ensure the semantic consistency between original and augmented data when generating diverse data.

We propose DRDA to generate augmentation data without introducing semantic perturbations. DRDA augments original data with multi-granularity segmentations, and pulls representations of one sentence closer with multi-view learning. Furthermore, we propose a dynamic selection technique to automatically choose an appropriate granularity in inference.

4.1 Multi-Granularity Segmentations

DRDA constructs symbolically diverse and semantically consistent augmentation data with multi-granularity segmentations. The point is that multi-granularity subword segmentation is a reversible process that completely retains semantic information, and is a deterministic process that always chooses the most probable and appropriate subword segmentation policy.

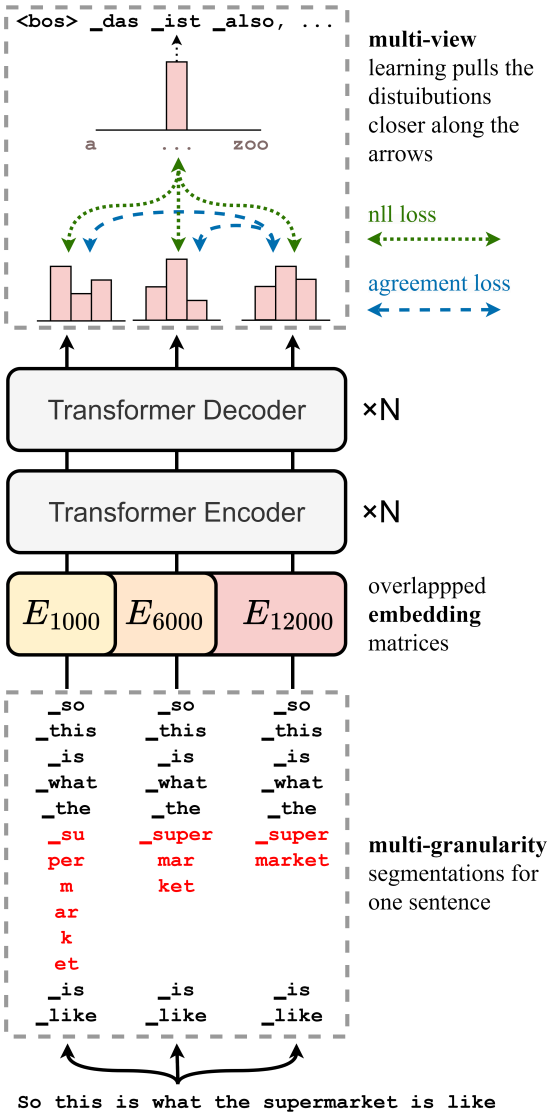

Figure 2: Illustration of the overall framework of DRDA. A source sentence is segmented into different granularities, and every generated token sequence will go through the model, obtaining a hypothesis distribution respectively. The agreement loss (blue segmented lines) will be computed between hypothesis distributions, and the negative likelihood loss (green dotted lines) will be computed between each distribution and the target.

Formally, given a prime vocabulary size p and a set of augmented vocabulary sizes $\{q_i\}_{i=1}^k$, for a source-target translation pair sample $(\mathbf{s},\mathbf{t})$, a prime source sequence $\mathbf{x}^{pri}$, a target sequence $\mathbf{y}$ and a set of augmented source sequences $\{\mathbf{x}^{aug_i}\}_{i=1}^k$ can be generated:

$$\mathbf{x}^{pri} = \arg\max_{\mathbf{x}\in V_p} P(\mathbf{x}|\mathbf{s}), \tag{3}$$

$$\mathbf{x}^{aug_i} = \arg\max_{\mathbf{x} \in V_{q_i}} P(\mathbf{x}|\mathbf{s}), \tag{4}$$

$$\mathbf{y} = \arg\max_{\mathbf{y}' \in V_p} P(\mathbf{y}'|\mathbf{t}). \tag{5}$$

Figure 2 depicts the model architecture and training loss on a English→Germany sample. Given $p = 12000$, $q_1 = 1000$, and $q_2 = 6000$, an English sentence is segmented with different vocabularies, generating three token sequences with different granularities.

Note that according to the greedy property of BPE, a short vocabulary is a prefix of a long vocabulary, as long as they are obtained from the same corpus. As a result, introducing different granularities with BPE will not lead to a larger vocabulary, thus avoiding an increase in parameter size. An example is shown in Figure 2, where three embedding matrices E_{12000}, E_{6000} and E_{1000} are overlapped, and a smaller embedding is a prefix of a larger embedding.

4.2 Multi-view Learning

Moreover, to make the translation model learn from different segmentation granularities, we utilize the multi-view learning loss function (Wang et al., 2021; Kambhatla et al., 2022b) and pull different representations closer together:

$$\begin{aligned}\mathcal{L} &= \underbrace{\mathcal{L}_{NLL}(P(\mathbf{y}|\mathbf{x}^{pri};\theta))}_{\text{prime source loss}} \\ &+ \underbrace{\frac{1}{k}\sum_{i=1}^{k}\mathcal{L}_{NLL}(P(\mathbf{y}|\mathbf{x}^{aug_i};\theta))}_{\text{augmented source loss}} \\ &+ \underbrace{\frac{\alpha}{k}\sum_{i=1}^{k}\mathcal{L}_{dist}(P(\mathbf{y}|\mathbf{x}^{pri};\theta), P(\mathbf{y}|\mathbf{x}^{aug_i};\theta))}_{\text{agreement loss}},\end{aligned} \tag{6}$$

where $\mathcal{L}_{NLL}$ is the negative likelihood loss in machine translation, $\mathcal{L}_{dist}$ is the symmetric Kullback-Leibler divergence (Kambhatla et al., 2022b).

The first two terms of Equation 6 (prime source loss and augmented source loss) compute the translation loss for source and augmented sentences respectively, and the third term (agreement loss) pulls the prediction distributions of different source inputs together.

As shown in Figure 2, output probability distributions for all granularities are used to compute the loss, where the blue segmented lines refer to the agreement loss between different granularities, and green dotted lines refer to the negative likelihood loss between the prediction and the target.

4.3 Dynamic Selection of Granularity in Inference

DRDA employs multiple segmentations in different granularities, so the selection of the granularity used in inference becomes a concern. To automatically choose a suitable vocabulary size when inferring, we also propose a simplified but granularity-focused version of n-best decoding (Kudo, 2018) to dynamically select the segmentation granularity in inferring step.

Given the set of all prime and augmented vocabulary sizes $\{p, q_1, q_2, \cdots, q_k\}$ and an input sentence $\mathbf{s}$, a series of $(\mathbf{x}, \mathbf{y})$ pairs can be generated, where each $(\mathbf{x}, \mathbf{y})$ pair represents a source-target token sequence pair in a certain granularity.

The estimated most probable segmentation and translation pair corresponds to the $(\mathbf{x}, \mathbf{y})$ pair that maximizes the following score:

$$score(\mathbf{x}, \mathbf{y}) = \log P(\mathbf{y}|\mathbf{x})/|\mathbf{y}|, \tag{7}$$

where $|\mathbf{y}|$ is the length of $\mathbf{y}$.

5 Experiments

We evaluate DRDA with translation tasks in different language pairs and translation directions to show its universal property regardless of language features. We also conduct experiments on extremely low resources and noisy scenarios to show the robustness of DRDA.[2]

5.1 Experimental Setup

	WMT En → De	IWSLT En ↔ (De, Fr, Zh, Es)	TED En ↔ Sk
train	4.5M	160k, 236k, 235k, 183k	61k
valid	3000	7283, 9487, 9428, 5593	2271
test	3003	6750, 1455, 1459, 1305	2445

Table 1: Overviews of datasets and corresponding sizes.

Datasets and preprocessing Our experiments are conducted on different datasets, as detailed in Table 1. We experiment on a low resource setting with IWSLT datasets, including IWSLT14 En↔De, En↔Es, and IWSLT17 En↔Zh, En↔Fr. We use

[2]Further setup details about dataset split, preprocessing, models, and evaluation are listed in Appendix A.

Model	IWSLT								WMT
	En→De	De→En	En→Fr	Fr→En	En→Zh	Zh→En	En→Es	Es→En	En→De
Transformer	29.03	35.26	37.57	37.29	22.38	21.29	39.92	41.86	27.08
DRDA	30.84‡	37.90‡	**38.77‡**	**38.55†**	**23.36†**	22.64†	41.99‡	43.90‡	27.41†
DRDA dyn.	**30.92‡**	**37.95‡**	38.75†	38.52†	23.32†	**22.90†**	**42.07‡**	**44.08‡**	**27.45†**

Table 2: BLEU on IWSLT and WMT. Statistical significance over Transformer is indicated by † ($p < 0.05$) and ‡ ($p < 0.001$). Significance is computed via bootstrapping (Koehn, 2004) using `compare-mt` (Neubig et al., 2019).

larger WMT14 En→De as a high-resource scenario dataset. The performance in extremely low resource scenarios is explored with the TED En↔Sk dataset. Following previous work (Vaswani et al., 2017), we lowercase words in IWSLT En↔De, while keeping other datasets cased.[3]

Models We build models on top of Transformer (Vaswani et al., 2017) with `Fairseq` toolkit (Ott et al., 2019). We use a Base Transformer model `transformer_wmt_en_de` for WMT, and `transformer_iwslt_de_en` for others.

Hyperparameters in training and inferring We use `sentencepiece` (Kudo and Richardson, 2018) to perform tokenization and BPE segmentation. The BPE encoding model is learned jointly on the source and target sides except for IWSLT En↔Zh. Unless otherwise stated, we use two vocabulary tables (on prime vocabulary and one augmented vocabulary), and their vocabulary sizes follow Table 3. Detailed analysis of the vocabulary sizes and the number of augmented vocabularies will be shown in Section 6.1. The weight of agreement loss α is set to 5 unless otherwise stated.

	WMT	IWSLT	TED
DRDA pri	32k	10k	8k
DRDA aug	16k	5k	4k
others	32k	10k	8k

Table 3: Prime and augmented vocabulary sizes used in DRDA, and vocabulary sizes used in other methods.

Evaluation We evaluate the performance of NMT systems using BLEU. To compare with previous work (Vaswani et al., 2017; Kambhatla et al., 2022b), we apply multi-bleu with `multi_bleu.perl` [4] for IWSLT En↔De, WMT En→De, and TED En↔Sk. For WMT En→De dataset, we additionally apply compound splitting[5]. All other datasets are evaluated with `SacreBLEU`[6].

5.2 Main Result

We present the results of DRDA on IWSLT and WMT translation tasks in Table 2. We can see that DRDA consistently outperforms the Transformer with a clear margin on all translation tasks. Moreover, models inferred with the dynamic granularity selection obtain a modest improvement in DRDA.

Model	En→De	De→En
Transformer	29.03	35.26
WordDrop	29.21	35.60
SwitchOut	29.00	35.90
RAML	29.70	35.99
DataDiverse	30.47	37.00
BPE-Drop	30.16	36.54
SubwordReg	29.46	36.14
R-Drop	30.45	37.40
MVR	30.44	37.47
CipherDAaug	30.65	37.60
DRDA	30.84‡	37.90‡
DRDA dyn.	**30.92‡**	**37.95‡**

Table 4: BLEU scores on IWSLT En↔De. Results of previous data augmentation (the second to the fifth models) are cited from literature which we share the same configuration with, as detailed in Appendix A.

Comparison between DRDA and other data augmentation and subword regularization methods on IWSLT are shown in Table 4. We use a range of augmentation and regularization methods for comparison. The augmentation methods include WordDrop (Zhang et al., 2020; Sennrich et al., 2016a),

[3] En, De, Fr, Zh, Es, Sk stand for English, German, French, Chinese, Spanish, and Slovak respectively.

[4] mosesdecoder/scripts/generic/multi-bleu.perl

[5] tensorflow/tensor2tensor/utils/get_ende_bleu.sh

[6] SacreBLEU signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.0`

SwitchOut (Wang et al., 2018b), RAML (Norouzi et al., 2016) and Data Diversification (Nguyen et al., 2020). The subword regularization methods include BPE-Drop (Provilkov et al., 2020) and Subword Regularization (Kudo, 2018). We also compare our method with others that adopt multi-view learning techniques, including R-Drop (Wu et al., 2021), MVR (Wang et al., 2021), and CipherDAug (Kambhatla et al., 2022b). DRDA yields greater improvement compared to others.

5.3 Extremely Low Resource Setting

TED En↔Sk task is challenging because of its extremely low resources (only 61k training sentence pairs). Several techniques have been adopted to improve the performance in low-resource NMT tasks like this, including data augmentation, multilingual translation, and transfer learning (Ranathunga et al., 2021). Neubig and Hu (2018) firstly propose similar language regularization to mix low-resource language with a lexically related high-resource language, combining transfer learning and multilingual translation. Several works continue to extend SRL and achieve high translation quality (Xia et al., 2019; Ko et al., 2021; Wang et al., 2018a).

Model	**En→Sk**	**Sk→En**
Transformer	20.82	28.97
LRL+HRL	-	32.07
CipherDAaug	24.61	32.62
DRDA	24.48	33.25
DRDA dyn.	**24.67**	**33.34**

Table 5: BLEU scores on TED En↔Sk. LRL+HRL method combines the original low-resource language pair with a high-resource related language Czech.

On this task, DRDA yields stronger improvements over baseline Transformer than other techniques with no requirement for external high resource languages, as shown in Table 5.

5.4 Robustness to Perturbations

We validate the robustness of DRDA on two noisy datasets. The first one is IWSLT De→En test set with synthetic perturbations. The perturbations are synthesized by traversing every character excluding space and punctuation in source sentences, and applying one of the operations with probability 0.01: (1) remove the character, (2) add a random character following the character, and (3) substitute the character with a random one. The second dataset is himl test set[7], which contains health information and scientific summaries and differs considerably from the IWSLT training set. Cross-domain datasets have different subword distributions, and the difference can be viewed as a natural noise. The results of the noisy test sets are shown in Table 6.

	original	synthetic	himl
Transformer	35.26	32.19	26.11
R-Drop	37.40	34.34	28.15
BPE-Drop	36.54	**35.00**	27.92
SubwordReg	36.14	34.55	27.63
DRDA	37.90	34.94	**28.80**
DRDA dyn.	**37.95**	34.98	28.78

Table 6: BLEU scores on original and noisy IWSLT De→En test set, and himl test set. Models are trained on the IWSLT De→En training set.

Along with these results, consistent improvement over Transformer and R-Drop is obtained by DRDA on both synthetically noisy and cross-domain datasets. DRDA significantly outperforms subword sampling methods (BPE-Drop and subword regularization) on natural noise datasets, but only obtains similar results with synthetic noise. We will discuss the reason in Section 6.2.

6 Analysis

In this section, we conduct analysis experiments to answer the following research questions (RQs) respectively:

- RQ1 (ablation satudies): How do the applied techniques and components affect model performance?
- RQ2: Does our approach really keeps semantic consistency between original and augmented data?
- RQ3: How does multi-granularity segmentation improve subword representations?
- RQ4: Why does multi-view learning help improve NMT models?

6.1 RQ1: Ablations

Choice of vocabulary sizes Here, we investigate the effects of pre-defined vocabulary sizes. As is mentioned in Section 5.1, we adopt one prime vocabulary and one augmented vocabulary. To find

[7] https://www.himl.eu/test-sets

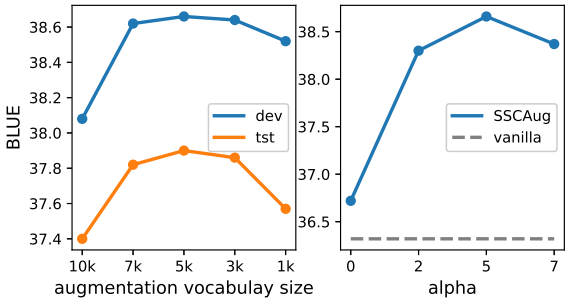

Figure 3: Ablations on IWSLT De→En over augmented vocabulary size (left) and agreement loss weight (right).

the optimal vocabulary sizes, we test {10k, 7k, 5k, 3k, 1k} for augmented vocabulary size when the prime size is 10k. Figure 3 verifies that, when the augmented vocabulary size is around 5k, the NMT model obtains the highest BLEU. The intuition is a huge difference in prime and augmented vocabulary sizes may corrupt the subword semantics, while a tiny difference may reduce the symbolic difference. A general recommendation in choosing vocabulary sizes is to use a proven suitable size for the prime vocabulary and set the augmented size to half the size of the prime vocabulary.

Weight of agreement loss As is shown in Figure 3, we find that agreement loss weight α significantly affects the performance of our method. Models obtain the highest BLEU score when $\alpha = 5$, and increasing or decreasing α causes a score drop up to 2 BLEU on the valid set. The model without agreement loss (i.e., $\alpha = 0$) still outperforms vanilla Transformer, validating the important role multi-granularity segmentation plays in DRDA.

{1k}	{6k}	{9k}	μ	σ
37.90	38.17	37.95	38.01	0.12
{1k,6k}	**{1k,9k}**	**{6k,9k}**	μ	σ
38.21	38.20	38.16	38.19	0.02

Table 7: BLEU scores on IWSLT De→En valid set when a 12k prime vocabulary is combined with different augmented vocabulary sets. μ and σ refer to the mean and standard deviation of BLEU scores when combined with one (top) or two (bottom) augmented vocabularies.

Number of augmented vocabularies Table 7 shows the effects of adding an extra augmented vocabulary with a prime vocabulary size of 12k on the valid set. When combined with two augmented vocabularies, the BLEU scores have a smaller deviation than combined with one. We can summarize that adding extra augmented vocabularies helps get a steady, comparable, and maybe slightly better result in the cost of an increase in training time.

6.2 RQ2: Semantic Consistency

Appendix B theoretically validates that our approaches generates more appropriate segmentations of a same sentence than other subword regularization methods. As a result, although both DRDA and subword regularization are reversible, DRDA is semantically more consistent because of the segmentation appropriateness.

To give an empirical insight of the semantical consistency, we analyze the nearest neighbors of subwords of different models (shown in Table 8), we find that vanilla Transformer and DRDA both exhibit semantics-based neighbors, where the embeddings of synonyms are similar. However, embeddings obtained in BPE-Drop tend to have high similarity with those they share a common sequence. Although this tendency can effectively alleviate vulnerability to misspelling, which explains the superiority subword regularization shows in synthetic noisy data in Section 5.4, it may introduce semantic error as well (treat "_go" as a synonym for "_good" in Table 8 for example), causing inaccuracy in machine translation.

	_good	
Transformer	DRDA	BPE-Drop
_great	_great	_great
_better	_big	_go
_nice	_nice	_bad
_bad	_bad	_god
_useful	_significant	_nice

Table 8: Top 5 nearest neighbors of subwords "_good" on IWSLT De→En.

The observation above indicates that DRDA introduces little semantic noise to augmentation data, and exhibits better semantic consistency.

6.3 RQ3: Effects on Subword Frequency

Here, we show that the mechanism of multi-granularity segmentation can be attributed to the increase in frequency of infrequent tokens.

NMT models with larger vocabulary sizes have larger atomic translation units, i.e., more coarse-grained subwords, so that they can better memorize one-to-many or many-to-one mappings and

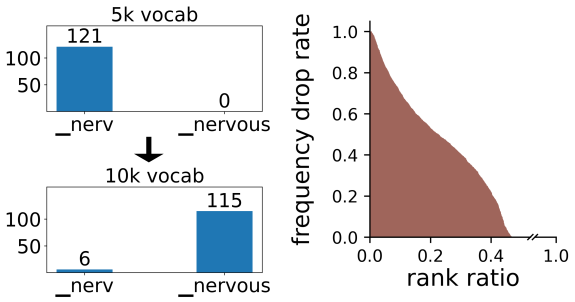

Figure 4: Most occurrences of "_nerv" are absorbed by "_nervous" when the vocabulary grows (left). The frequency drop rate of "_nerv" is $(121-6)/121 = 0.95$. The right figure shows all frequency drop rates on IWSLT En→De sorted in descending order.

resolve translation ambiguity (Koehn, 2009). However, fine-grained subwords may suffer from a frequency drop when the vocabulary size grows. Figure 4 shows that most occurrences of "_nerv" are absorbed by "_nervous" when the vocabulary grows, making it more difficult for the NMT model to obtain a precise representation of other inflection forms like "_nervy", "_nervier" and "_nervine". More generally, the frequency drop is common on IWSLT En→De (results on more datasets are shown in Appendix D), where about 50% of subwords appeared in 5k vocabulary suffer from a frequency drop when the vocabulary grows to 10k, as Figure 4 shows.

In DRDA, by taking both small and large vocabulary sizes simultaneously, infrequent tokens occur more frequently so that subwords like "_nerv" can be trained in adequate contexts as well.

6.4 RQ4: Multi-view Techniques and Subword Semantic Composition

Multi-view learning pulls representations in different granularities together. To investigate the effects of multi-view techniques, we propose a task to find out how the coarse-grained and fine-grained representations of the same word are drawn closer.

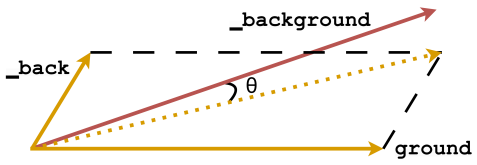

Figure 5: The similarity between the fine- and coarse-grained representations is computed by $\cos\theta$.

The process is illustrated with an example in Figure 5, and the formal definition of the task is shown in Appendix C. We take a coarse-grained subword ("_background") and its corresponding fine-grained subword sequence ("_back", "ground"), then compute the cosine similarity between the former embedding and the sum of the latter embeddings. The similarity indicates the extent to which the fine-grained and coarse-grained representations are brought closer together.

We enumerate all the coarse-grained and fine-grained representation pairs, and average all their cosine similarity scores. The results are shown in Table 9. As expected, DRDA with proper agreement loss ($\alpha = 5$) obtains a higher average similarity than other data augmentation approaches.

Model	_back, ground	_plat, form	_feed, back	avg
Transformer	0.22	0.28	0.31	0.24
R-Drop	0.41	0.36	0.39	0.35
BPE-Drop	0.54	0.62	0.66	0.46
DRDA $\alpha = 0$	0.48	0.50	0.66	0.50
DRDA $\alpha = 2$	0.71	0.67	0.77	0.67
DRDA $\alpha = 5$	0.78	**0.79**	**0.85**	**0.77**
DRDA $\alpha = 7$	**0.81**	0.74	0.82	0.75

Table 9: Similarities between coarse- and fine-grained representations for the same word (e.g., "_background" vs. "_back"+"ground"). **avg** refers to the average similarities of all words on IWSLT En→De.

Computing the similarities between representations in multiple granularities is a subword level composition (SSC) tasks (Mitchell and Lapata, 2008, 2009; Turney, 2014). We can conclude that multi-view techniques help DRDA models improve the SSC understanding, thus obtaining better robustness to perturbations (Provilkov et al., 2020).

7 Conclusion

In this paper, we identify the semantic inconsistency caused by irreversible operations or probabilistic segmentations, and propose a deterministic reversible data augmentation consisting of multi-granularity segmentation and multi-view learning to ensure the consistency when generating diverse data. Experiments demonstrate the superiority of our proposed DRDA over previous data augmentation and subword regularization in terms of translation accuracy and robustness. We also offer a combination of empirical and theoretical verification of semantic consistency, and insightful analyses about multi-granularity and multi-view techniques.

Limitations

High resource scenarios As other data augmentation techniques, our proposed DRDA appears to be less effective in high-resource scenarios. The analysis in Section 6.3 offers one explanation to this phenomenon that, the frequency drop becomes less sharp when the data size grows, thus resulting in lower effectiveness of data augmentation. Considering this phenomenon, a better application approach of data augmentation on high-resource scenarios can be designed, by locating the rare subwords of a specific domain in a model trained on large general corpus and continuing training with the augmentation data. We leave this investigation as a direction for future research.

Application scope As a foundation process in NLP, segmentation is applied in various tasks, including language modeling, named entity recognition, and numerous others. As a result, segmentation based data augmentation techniques including DRDA can be applied to a wide range of tasks. One limitation of this study is its exclusive application of DRDA to machine translation, which restricts the ability to validate and compare its effectiveness across other tasks.

References

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.

Huadong Chen, Shujian Huang, David Chiang, Xinyu Dai, and Jiajun Chen. 2018. Combining character and word information in neural machine translation using a multi-level attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1284–1293.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5961–5970.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974.

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.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Bi-simcut: A simple strategy for boosting neural machine translation. *arXiv preprint arXiv:2206.02368*.

Yingqiang Gao, Nikola I Nikolov, Yuhuang Hu, and Richard HR Hahnloser. 2020. Character-level translation with self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1591–1604.

Jie Hao, Xing Wang, Shuming Shi, Jinfeng Zhang, and Zhaopeng Tu. 2019. Multi-granularity self-attention for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 887–897.

Nishant Kambhatla, Logan Born, and Anoop Sarkar. 2022a. Auxiliary subword segmentations as related languages for low resource multilingual translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 131–140.

Nishant Kambhatla, Logan Born, and Anoop Sarkar. 2022b. Cipherdaug: Ciphertext based data augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 201–218.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource nmt models to translate low-resource related languages without parallel data. 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 802–812.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

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

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Bei Li, Tong Zheng, Yi Jing, Chengbo Jiao, Tong Xiao, and Jingbo Zhu. 2022. Learning multiscale transformer models for sequence generation. In *International Conference on Machine Learning*, pages 13225–13241. PMLR.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuan-Jing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.

Jeff Mitchell and Mirella Lapata. 2009. Language models based on semantic composition. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 430–439.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. *Advances in Neural Information Processing Systems*, 33:10018–10029.

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.

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.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. Bpe-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

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 (Volume 1: Long Papers)*, 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 arXiv:2009.13818*.

Peter D Turney. 2014. Semantic composition and decomposition: From recognition to generation. *arXiv preprint arXiv:1405.7908*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2018a. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018b. 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.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2021. Multi-view subword regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Lijun Wu, Shufang Xie, Yingce Xia, Yang Fan, Jian-Huang Lai, Tao Qin, and Tieyan Liu. 2020. Sequence generation with mixed representations. In *International Conference on Machine Learning*, pages 10388–10398. PMLR.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

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.

A Implementation Details

A.1 Datasets and Preprocessing

We perform minimum preprocessing and cleaning steps to raw data.

- For IWSLT En↔De, following previous works, data is obtained with `fairseq` scripts[8], which performs `clean-corpus-n` [9] with ratio $= 1.5$, $\min = 1$ and $\max = 175$.
- For other IWSLT datasets, we extract titles, descriptions and main texts for training, and main texts only for validating and testing. There is no extra cleanup operation performed. IWSLT14 En↔Es dataset concatenates `dev2010`, `tst2010`, `tst2011` and `tst2012` as development set, uses `tst2015` as test set. IWSLT17 En↔Fr and En↔Zh datasets concatenate `dev2010`, `tst2010`, `tst2011`, `tst2012`, `tst2013`, `tst2014` and `tst2015` as development set, use `tst2017` as test set.
- We use `t2t-datagen` [10] script to generate WMT data, and performs `clean-corpus-n` with $\min = 1$ and $\max = 80$, removing about 1% training sentence pairs. Following previous works, we validate on `newstest2013` and test on `newstest2014`.
- The TED datasets are obtained using scripts from the official repository (Qi et al., 2018). We additionally remove the encoder language token "`__sk__`" to accommodate bilingual NMT.

A.2 Models and Hyperparameters

Smaller datasets are trained with model `transformer_iwslt_de_en`, and WMT dataset is trained with model `transformer_wmt_en_de`. The corresponding config is shown in Table 10.

	small	base
encoder layer	6	6
decoder layer	6	6
attention head	4	8
embedding size	512	512
feed-forward size	1024	2048
learning rate	$6e-4$	$5e-4$
lr schedule	inverse sqrt	inverse sqrt
optimizer	adam	adam
adam betas	(0.9, 0.98)	(0.9, 0.98)
clip norm	0.1	-
warm updates	8000	4000
network dropout	0.3	0.1
attention dropout	0.1	0.1
weight decay	$1e-4$	0
label smoothing	0.1	0.1
words per batch	about 17k	about 380k
beam	5	4
length penalty	1.0	0.6

Table 10: Model configuration of `transformer_iwslt_de_en` (small) and `transformer_wmt_en_de` (base).

Note that DRDA, R-Drop, CipherDAug and some other approaches may double the input texts, but we constrain the tokens number forwarded to the model in a batch according to the "words per batch" hyperparameter, which means the numbers

[8] fairseq/example/translation//prepare_iwslt14.sh

[9] mosesdecoder/scripts/training/clean-corpus-n.perl

[10] tensorflow/tensor2tensor/bin/t2t-datagen

of sentences in a batch of these approaches are rough halved.

A.3 Computational Cost

Total training duration and GPU used in DRDA experiments are listed in Table 11.

	WMT	IWSLT	TED
GPU	2× RTX 3090	4× TITAN Xp	2× TITAN X (Pascal)
time	12 days	2 days	10 hours
steps	100k	200k	35k

Table 11: Computational cost of WMT, IWSLT and TED experiments.

A.4 Baseline Implementation

We reimplement those models with high relevance with our method, including vanilla Transformer, BPE-Drop, subword regularization, R-Drop, MVR and CipherDAaug. These models except for Transformer use either segmentation-related techniques or multi-view techniques. Important details of our implementation are listed below:

- BPE-Drop and subword regularization are implemented using `sentencepiece`. In encoding, we set $\alpha = 0.1$ and $\alpha = 0.2$ for BPE-Drop and subword regularization respectively, and $\text{nbest\_size} = -1$ for both. Results of subword regularization are obtained without n-best decoding (Kudo, 2018).
- We use the task and loss module from the official open-source repository [11] to implement R-Drop. Weight α is set to 5.
- In MVR implementation, we adopt the same subword regularization hyper-parameters to BPE-Drop, and the agreement loss weight is set to be 5.
- CipherDAaug models are reimplemented on top of the official open-source code[12]. Following their instructions, we adopt 2 keys, and set agreement loss weight $\beta = 5$.

For the traditional data augmentation methods (WordDrop, SwitchOut, RAML, DataDiverse) with which DRDA shares a relatively low similarity, results are cited from Kambhatla et al. (2022b). We share exactly the same model architecture and hyperparameters with Kambhatla et al. (2022b), and we successfully reimplemented their main model with similar results, so we find it reliable to cite from.

We report the performance of LRL+HRL from the corresponding literature (Xia et al., 2019).

[11] https://github.com/dropreg/R-Drop

[12] https://github.com/protonish/cipherdaug-nmt

B Theoretical Discussion of Consistency

In this section, we discuss what is semantic consistency, and give a theoretical analysis about why DRDA is more semantically consistent.

It is clear that previous data augmentation methods that adopt irreversible operations result in semantic loss, which will inevitably do damage to the consistency between original and augmented data. DRDA is superior to these methods in terms of preservation of the original meanings, because it is based on reversible segmentation to generate diversity.

However, it is more tricky to prove that subword regularization methods (Kudo, 2018; Provilkov et al., 2020), which is also based on reversible segmentation, leads to more inconsistency that DRDA. To show the superiority of DRDA in consistency over subword regularization, we review the difference of the two in sampling segmentation in Section 3 and 4.1:

$$\mathbf{x}^i_{DRDA} = \arg\max_{\mathbf{x}} P_{seg}(\mathbf{x}|\mathbf{s}; p_i), \tag{8}$$

$$\mathbf{x}_{SR} \sim \arg\max_{\mathbf{x}} P_{seg}(\mathbf{x}|\mathbf{s}; p), \tag{9}$$

where $\mathbf{x}^i_{DRDA}$ is a representation in certain granularity of source sentence s in DRDA, $\mathbf{x}_{SR}$ is the representation in subword regularization, p_i and p are vocabulary sizes.

$\arg\max_{\mathbf{x}} P_{seg}(\mathbf{x}|\mathbf{s}; p)$ can be interpreted as the difficulty of segmenting s with a certain vocabulary size p. We can assume that the difficulty of segmenting a sentence is an inherent property of sentences, independent of vocabulary sizes:

$$\arg\max_{\mathbf{x}} P_{seg}(\mathbf{x}|\mathbf{s}) = \arg\max_{\mathbf{x}} P_{seg}(\mathbf{x}|\mathbf{s}; p), \tag{10}$$

where $p \in \mathbb{N}$ is any pre-defined vocabulary size.

Then, because of the deterministic argmax operation in DRDA and the random sampling operation in subword regularization, the following inequality holds:

$$P_{seg}(\mathbf{x}_{DRDA}|\mathbf{s}) \geq P_{seg}(\mathbf{x}_{SR}|\mathbf{s}). \tag{11}$$

Equation 11 validates that our approaches generates more appropriate segmentations of a same

sentence that other subword regularization methods. As a result, although both DRDA and subword regularization are reversible, DRDA is semantically more consistent because of the segmentation appropriateness.

C Process of Subword Semantic Composition Task

Let $a \circ b$ be a compound token concatenated by a and b, with their corresponding embedding $\mathbf{e}_{a\circ b}$, $\mathbf{e}_a$ and $\mathbf{e}_b$, the SSC understanding ability is scored by the similarity between $\mathbf{e}_{a\circ b}$ and $\mathbf{e}_a + \mathbf{e}_b$:

$$\mathrm{SIM}(\mathbf{e}_{a\circ b}, \mathbf{e}_a + \mathbf{e}_b) = \frac{\mathbf{e}_{a\circ b} \cdot (\mathbf{e}_a + \mathbf{e}_b)}{\|\mathbf{e}_{a\circ b}\| \cdot \|\mathbf{e}_a + \mathbf{e}_b\|}. \quad (12)$$

To numerically evaluate the superiority of a model in understanding SSC, we average semantic composition similarities of all subwords except characters and special tokens (such as <unk>):

$$\overline{\mathrm{SIM}} = \frac{1}{|\widetilde{V}|} \sum_{a,b,a\circ b \in V} \mathrm{SIM}(\mathbf{e}_{a\circ b}, \mathbf{e}_a + \mathbf{e}_b), \quad (13)$$

where $\widetilde{V}$ is a set of all subwords except characters and special tokens, and V is a set of all subwords.

It should be noted that the models listed in Section 6.4 share the same V, so that comparing the scores completely makes sense.

D More Studies

D.1 Subword Nearest Neighbors

_good		ood	
DRDA	BPE-Drop	DRDA	BPE-Drop
_great	_great	oods	oo
_big	_go	_penguins	oods
_nice	_bad	ago	ook
_bad	_god	_birthday	od
_significant	_nice	wow	_food
_better	_useful	ghter	_good
_huge	ood	_entrop	wood
_happy	_goods	anced	ool
_useful	_better	gal	_blood
_healty	_big	_astero	idung
_photograph		**_photographs**	
DRDA	BPE-Drop	DRDA	BPE-Drop
_photo	_phot	_pictures	_photos
_photos	_photographs	_photos	_pictures
_photographs	_photo	_images	_photograph
_picture	_fotograf	_movies	_phot
_pattern	_picture	_structures	ć
_digit	_ph	_chemicals	_images
_penguins	_graph	_statistics	t’
_tremend	ograph	_maps	ē
_pictures	t’	_scene	_
_doctors	_photos	_spectrum	©

Table 12: Top 10 nearest neighbors of example subwords.

Following Provilkov et al. (2020), we study the closest neighbors of word embedding learned in BPE-Drop and DRDA. Several examples are shown in Table 12.

We can find that in the morphology of words, BPE-Drop tends to bring two subwords sharing a common sequence together ("_good" and "ood" for example), while DRDA has no such behavior. On one hand, the tendency to pull similarly spelled words closer can effectively help NMT model overcome the perturbation of misspelling, as shown in previous experiments. On the other, it can introduce unreasonable noise as well, since similarly spelled subwords are not necessarily semantically related words ("_good" and "_go" for example).

D.2 Effects of Granularity Selection

Our experiments have shown that the dynamic selection of segmentation granularity yields a modest improvement in BLUE scores. Here, we investigate the mechanism and potential of this method.

	prime	augmented	dynamic	oracle
En → De	30.84	30.85	30.92	31.36
De → En	37.90	37.88	37.95	38.46
En → Fr	38.77	38.57	38.75	40.15
Fr → En	38.55	38.44	38.52	39.59
En → Zh	23.36	23.32	23.32	23.37
Zh → En	22.64	22.84	22.90	23.62
En → Es	41.99	42.07	42.07	42.64
Es → En	43.90	43.97	44.08	45.30

Table 13: BLEU scores on IWSLT tasks.

We define an oracle granularity selection model, whose translation result corresponds to the one with the highest sentence-level BLEU score among the results generated by source sequences with different granularities. The results of models with 5k augmented size and 10k prime size on IWSLT translation tasks are shown in Table 13.

It can be summarized from the results that the selection of input granularities has considerable potential (up to 1.7 BLEU) in improving the translation, and our approach obtains an improvement of up to 0.24 BLEU. In the future, a better re-ranking approach can be adopted to build a selection model closer to the oracle model.

D.3 Frequency Drops

More examples of frequency drop on WMT, IWSLT and TED are shown in Figure 6. Among

these results, all datasets suffer from a similar frequency drop regardless of their language directions and sizes. The vocabulary size grows from 16k to 32k for WMT, from 5k to 10k for IWSLT and from 4k to 8k for TED.

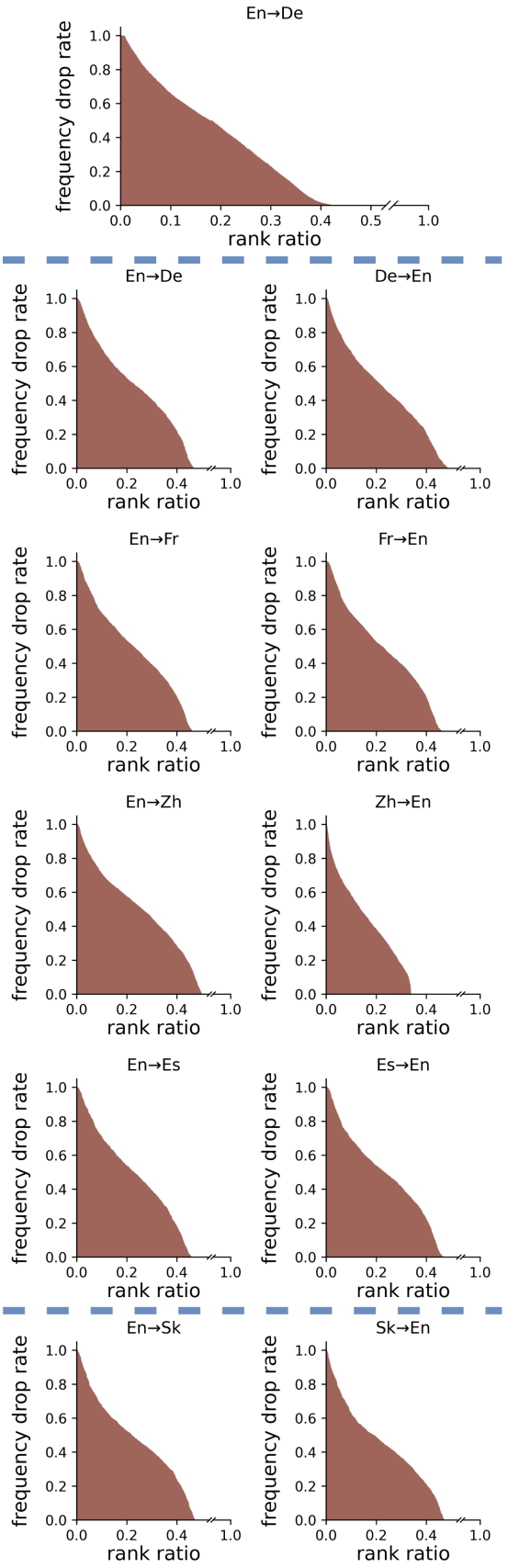

Figure 6: Subwords frequency drop rates on WMT (top), IWSLT (middle), and TED (bottom).