Sentence-Level or Token-Level? A Comprehensive Study on Knowledge Distillation

Jingxuan Wei[1,2*], **Linzhuang Sun**[1,2], **Yichong Leng**[3], **Xu Tan**[4*], **Bihui Yu**[1,2], **Ruifeng Guo**[1,2]
[1]Shenyang Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]University of Science and Technology of China
[4]Independent Researcher
weijingxuan20@mails.ucas.edu.cn, tanxu2012@gmail.com

Abstract

Knowledge distillation, transferring knowledge from a teacher model to a student model, has emerged as a powerful technique in neural machine translation for compressing models or simplifying training targets. Knowledge distillation encompasses two primary methods: sentence-level distillation and token-level distillation. In sentence-level distillation, the student model is trained to align with the output of the teacher model, which can alleviate the training difficulty and give student model a comprehensive understanding of global structure. Differently, token-level distillation requires the student model to learn the output distribution of the teacher model, facilitating a more fine-grained transfer of knowledge. Studies have revealed divergent performances between sentence-level and token-level distillation across different scenarios, leading to the confusion on the empirical selection of knowledge distillation methods. In this study, we argue that token-level distillation, with its more complex objective (i.e., distribution), is better suited for "simple" scenarios, while sentence-level distillation excels in "complex" scenarios. To substantiate our hypothesis, we systematically analyze the performance of distillation methods by varying the model size of student models, the complexity of text, and the difficulty of decoding procedure. While our experimental results validate our hypothesis, defining the complexity level of a given scenario remains a challenging task. So we further introduce a novel hybrid method that combines token-level and sentence-level distillation through a gating mechanism, aiming to leverage the advantages of both individual methods. Experiments demonstrate that the hybrid method surpasses the performance of token-level or sentence-level distillation methods and the previous works by a margin, demonstrating the effectiveness of the proposed hybrid method.

1 Introduction

Knowledge distillation, as a fundamental technique for model compression and knowledge transfer in deep neural networks, has wide application in the field of neural machine translation (NMT) [Hinton *et al.*, 2015; Gou *et al.*, 2021]. Knowledge distillation involves transferring knowledge from a larger, cumbersome model to a smaller, more efficient one, serving purposes such as compressing machine translation models and simplifying training targets for non-autoregressive models [Phuong and Lampert, 2019; Liu *et al.*, 2020; Wang and Yoon, 2021; Xiao *et al.*, 2023].

Given the variance in training targets, knowledge distillation in NMT can be divided into two main categories: sentence-level knowledge distillation and token-level knowledge distillation. Sentence-level knowledge distillation mainly focuses on simplifying the training target to improve the translation accuracy [Gajbhiye *et al.*, 2021; Yang *et al.*, 2022a]. Specifically, given a source and target sentence pair, sentence-level distillation firstly feeds the source sentence into the teacher model to generate a pseudo target sentence, then the pseudo target sentence is leveraged as the training target of student model. Compared with the origin target sentence, the distribution of pseudo target sentence is simpler, and thus easier to learn for student model [Kim and Rush, 2016; Zhang *et al.*, 2019; Tang *et al.*, 2019; Tan *et al.*, 2022].

In contrast, token-level knowledge distillation focuses on enhancing translation quality by a finer granularity [Kim and Rush, 2016; Mun'im *et al.*, 2019]. Different with sentence-level knowledge distillation which only leverages the output sentence of teacher model, token-level knowledge distillation further uses the token distribution in the output sentence. The student model is trained to output a similar distribution with the teacher model on every token, which helps the student model learn detail knowledge on token difference and be more suitable for texts with high lexical diversity [Wang *et al.*, 2020].

However, empirical studies have revealed divergent performances between sentence-level and token-level distillation across different scenarios. Specifically, while some scenarios benefit more from the global structure and semantic consistency provided by sentence-level distillation [Kim and Rush, 2016; Chen *et al.*, 2020; Xu *et al.*, 2021b; Lei *et al.*, 2022; Mhamdi *et al.*, 2023], other scenarios require the fine-grained knowledge transfer that token-level distillation offers [Liao *et al.*, 2020; Tang *et al.*, 2021; Li *et al.*, 2021; Ma *et al.*, 2023]. This variation in performance has led to confusion regarding the empirical selection of knowledge distilla-

tion methods. In this study, we conduct analytical experiments to explore the general suitable scenario of two knowledge distillation methods. Given that the training target of sentence-level distillation (simplified sentence by teacher model) is easier than that of the token-level distillation (detailed token distribution of teacher model). We hypothesize that sentence-level distillation is suitable for "complex" scenarios and the token-level distillation is suitable for "simple" scenarios.

We define the "complex" or "simple" scenarios from three perspectives: 1) model size of student model, as the student model becomes small, it is harder for the student model to learn the knowledge, and thus the scenario become more complex; 2) the complexity of text, more complex text will make the learning procedure of student model harder; 3) the difficulty of decoding, which is determined by the amount of auxiliary information available during decoding. The more auxiliary information available, the simpler the decoding process becomes. Experiments on the above three perspectives consistently verify our hypothesis, showing that the token-level distillation performs better in simple scenarios with the sentence-level distillation is better for complex scenarios.

Although the analytical experiments provide deep understanding and reveal the general suitable scenarios for two distillation method, how to empirically define the complexity of a machine translation task is challenging. To address this challenge, we further explore the hybridization of two distillation methods, aiming at taking the advantage of both the distillation methods to enhance overall translation accuracy. We propose a dynamic gating mechanism that adaptively balances the learning process between sentence-level and token-level distillation. Specifically, the student model is trained to learn both the pseudo target sentence distribution for global coherence and the detailed token distribution from the teacher model for lexical precision, with the gating mechanism dynamically adjusting the emphasis based on the evolving learning context and model performance.

The contributions of this paper are summarized as follows:

- We conduct experiments to discover the optimal use of sentence-level and token-level distillation, uncovering that the sentence-level distillation excels in simpler scenarios, whereas the token-level distillation is more effective in complex ones.
- We propose a hybrid method of sentence-level and token-level distillation, showing enhanced performance over single distillation methods and baseline models.

2 Related Work

Knowledge distillation (KD) is widely applied in the field of neural machine translation (NMT) to enhance the efficiency and performance of translation models [Hinton *et al.*, 2015; Xu *et al.*, 2021b; Chen *et al.*, 2020; Gou *et al.*, 2021; Zhang *et al.*, 2022]. Recently, knowledge distillation is applied in multilingual NMT [Tan *et al.*, 2019] to assist models in mastering multiple languages within a single framework. A multi-agent learning framework [Liao *et al.*, 2020] is utilized to investigate how sentence-level and token-level distillation can work together synergistically. PMGT [Ding *et al.*, 2021]enhances phrase translation accuracy and model reordering capability by progressively increasing the granularity of training data from words to sentences. ProKD [Ge *et al.*, 2023] demonstrates the use of high-resource language teacher models to enhance translation performance in low-resource languages through cross-lingual knowledge distillation. Despite the emergence of various distillation models, knowledge distillation in NMT, from the perspective of training targets, can be primarily divided into two categories: sentence-level knowledge distillation and token-level knowledge distillation.

2.1 Token-Level Knowledge Distillation

Token-level knowledge distillation in neural machine translation (NMT) primarily focuses on enhancing the translation accuracy of individual words or phrases [He *et al.*, 2021; Gou *et al.*, 2021; Wang and Yoon, 2021]. This approach is explored in various studies to improve specific aspects of translation quality. For example, token-level ensemble distillation for grapheme-to-phoneme conversion [Sun *et al.*, 2019] can enhance the phonetic translation accuracy. Additionally, a selective knowledge distillation method [Wang *et al.*, 2021] aims at optimizing the word-level distillation loss and the standard prediction loss. The raw data exposure model [Ding *et al.*, 2020] reduces lexical choice errors in low-frequency words by exposing NAT models to raw data, enhancing translation accuracy. SKD [Sun *et al.*, 2020] investigates knowledge distillation in the context of multilingual unsupervised NMT, while kNN-KD [Yang *et al.*, 2022b] examines the effects of nearest-neighbor knowledge distillation on translation accuracy. Furthermore, the token-level self-evolution training [Peng *et al.*, 2023] method dynamically identifies and focuses on under-explored tokens to improve lexical accuracy, generation diversity, and model generalization. The concept of knowledge distillation via token-level relationship graphs [Zhang *et al.*, 2024] offers a novel perspective on leveraging relational data for distillation, further contributing to the advancement of the token-level knowledge distillation in NMT.

2.2 Sentence-Level Knowledge Distillation

Sentence-level knowledge distillation in neural machine translation (NMT) focuses on reducing the training difficulty of student model, particularly useful in capturing the semantics of whole sentences or long sequences [Kim and Rush, 2016; Ren *et al.*, 2019; Stahlberg, 2020]. For examples, ensemble distillation method [Freitag *et al.*, 2017] is proposed to effectively combine multiple model outputs to improve the handling of complex sentence structures. The scope of sentence-level distillation techniques is further expanded with the help of perturbed length-aware position encoding in non-autoregressive neural machine translation [Oka *et al.*, 2021]. DDRS [Shao *et al.*, 2022] introduces diversified distillation and reference selection strategies to improve the accuracy of sentence-level distillation. Sentence-level distillation is also employed for simultaneous machine translation to address the challenges of real-time translation [Deng *et al.*, 2023].

Several studies have provided insights to better understand the knowledge distillation. For instance, NAT [Zhou *et al.*, 2020] delves into why knowledge distillation is effective in non-autoregressive machine translation (NAT), uncovering

Dataset	Teacher Size	Student Size	BLEU Score			
			Teacher Results	Token-level	Sentence-level	△
IWSLT14 de→en	38M	3M	34.80	30.50	31.09	-0.59
		9M		34.12	34.20	-0.08
		38M		36.09	34.84	1.25
		111M		36.40	34.87	1.53
IWSLT13 en→fr	52M	7M	44.10	39.63	41.94	-2.31
		12M		42.42	43.48	-1.06
		52M		44.82	44.43	0.39
		140M		44.87	44.26	0.61
WMT14 en→de	83M	28M	27.35	23.89	25.17	-1.28
		83M		26.49	26.77	-0.28
		112M		26.73	26.68	0.05
		146M		26.66	26.56	0.10
IWSLT17 ar→en	47M	13M	31.19	28.66	30.21	-1.55
		24M		29.02	30.52	-1.50
		47M		32.18	31.15	1.03
		84M		32.37	31.33	1.04

Table 1: Impact of model size on knowledge distillation across datasets. The △ column represents the difference between token-level and sentence-level BLEU scores. Positive values suggest that the token-level distillation has a higher BLEU score than the sentence-level distillation.

the impact of text complexity on NAT. However, this study does not explore how text complexity affects token-level and sentence-level distillation. HKD [Lee *et al.*, 2022] investigates the question of "when to distill such knowledge". It proposes a gate knowledge distillation scheme, where the teacher model serves not only as a knowledge provider but also as a calibration measurement, allowing for a switch between learning from the teacher model and training the student. This work also investigates both token-level and sentence-level distillation in teacher model. However, it treats them as separate strategies with independent token-level and sentence-level gates and fails to combine these two approaches. Our work explores the general suitable scenario of knowledge distillation for both token-level and sentence-level perspectives, hypothesizing that token-level distillation is better suited for 'simple' scenarios, while sentence-level distillation excels in 'complex' scenarios. Furthermore, we propose a hybrid method that combines token-level and sentence-level distillation through a gating mechanism, aiming to alleviate the empirical confusion on selecting the distillation methods.

3 Comprehensive Analysis of Knowledge Distillation

This section presents a detailed analysis of knowledge distillation within neural machine translation (NMT), focusing on the empirical evaluation of token-level versus sentence-level distillation in varied scenarios. This analysis aligns with our hypothesis outlined in Section 1: that sentence-level distillation is more adept in 'complex' scenarios, while token-level distillation excels in 'simple' scenarios. We define the complexity from three perspectives:

1) Model size of the student model: The scenarios become more complex when the model size of student model become smaller, since the student model need to compress the knowledge of teacher model into a model with limited capacity.

2) Complexity of the text: Datasets with more complex text, characterized by intricate sentence structures and diverse vocabulary, present more challenging learning environments for the student model.

3) Difficulty of decoding: The decoding difficulty is determined by the amount of ground truth or auxiliary information available during decoding. Scenarios where the decoder receives more ground truth or auxiliary information are considered simpler, as this additional information not only simplifies the decoding process by providing clearer guidance and reducing ambiguity, but also helps in avoiding the accumulation of errors during the decoding procedure.

In the following subsection, we firstly introduce the dataset and configuration used in the analysis experiments, then we verify our hypothesis from the above three perspectives.

3.1 Dataset and Configuration

For the experiments, we select four datasets to cover a range of complexities and linguistic characteristics: IWSLT13 English→French (en→fr), IWSLT14 German→English (de→en), WMT14 English→German (en→de), and IWSLT17 Arabic→English (ar→en). Each dataset offers a unique combination of bilingual sentence pairs and complexity levels: 200k for IWSLT13 en→fr, 153k for IWSLT14 de→en, 4.5M for WMT14 en→de, and 231k for IWSLT17 ar→en.

We apply byte-pair encoding (BPE) with subword-nmt toolkit[1] to all sentences in these datasets for tokenization. The

[1]https://github.com/rsennrich/subword-nmt

Dataset	Stud Size	Noise	BLEU Score				
			Token	Sentence	△	△ Rate (T)	△ Rate (S)
IWSLT14 de→en	38M	Orig	36.09	34.84	1.25	-	-
		Mod	34.31	33.68	0.63	-4.93%	-3.33%
		High	32.71	33.26	-0.55	-9.37%	-4.54%
IWSLT13 en→fr	18M	Orig	44.56	43.95	0.61	-	-
		Mod	42.89	42.50	0.39	-3.75%	-3.30%
		High	41.11	42.53	-1.42	-7.74%	-3.23%
WMT14 en→de	112M	Orig	26.73	26.68	0.05	-	-
		Mod	25.03	25.47	-0.44	-6.36%	-4.54%
		High	24.49	25.35	-0.86	-8.38%	-4.99%
IWSLT17 ar→en	47M	Orig	32.18	31.15	1.03	-	-
		Mod	30.24	30.15	0.09	-6.03%	-3.21%
		High	27.90	28.23	-0.33	-13.30%	-9.37%

Table 2: Impact of text complexity on knowledge distillation across datasets. The △ column represents the difference between token-level and sentence-level BLEU scores. The △ Rate (T) and △ Rate (S) columns represent the percentage decrease in BLEU scores from the original to moderate and high noise levels for token-level and sentence-level respectively.

vocabulary size is 32K. The experiments are conducted using the Fairseq[2] framework.

3.2 Impact of Model Size

In this subsection, we explore the impact of student model size on the effectiveness of token-level and sentence-level distillation. We adjust the size of the student model following the model size reduction approach in [Zhou *et al.*, 2020] to observe the impact of model size on knowledge distillation across different datasets. The results are shown in Table 1.

Analysis of Results and Summary

Our comprehensive analysis, as detailed in Table 1, reveals a clear relationship between the student model's size and the effectiveness of knowledge distillation methods. Across all datasets, we observed a consistent trend: as the model size increases, both token-level and sentence-level distillation methods show improvement in BLEU scores. This improvement is particularly notable in the transition from small to medium-sized models. For instance, in the IWSLT14 de→en dataset, a significant leap in performance is observed when the model size was increased from 3M to 9M parameters. However, beyond a certain threshold, such as 38M parameters in this dataset, the rate of improvement begins to plateau, indicating diminishing returns with further increases in size.

Interestingly, a critical point of inversion is observed where the advantage shifts from sentence-level to token-level distillation as the model size increases. In smaller models, sentence-level distillation tends to outperform token-level, aligning with our hypothesis that it is more suitable for complex scenarios where model size is limited. As the size increases, token-level distillation begins to show a relative advantage, suggesting its effectiveness in simpler scenarios with larger model capacities.

This trend suggests that while larger models can benefit from both distillation methods, there is an optimal range of model size where the gains are most substantial. Beyond this range, the additional complexity of larger models does not translate into proportional improvements in distillation performance. In practical terms, this implies that for scenarios prioritizing model compression, such as deploying NMT systems on resource-constrained devices, sentence-level distillation is more suitable due to its effectiveness in smaller models. Conversely, in scenarios where the focus is on maximizing translation accuracy, such as in server-based applications with fewer computational constraints or competition scenario [Farinha *et al.*, 2022; Blain *et al.*, 2023], token-level distillation becomes increasingly advantageous as model size grows.

3.3 Impact of Text Complexity

In this subsection, we investigate the impact of text complexity, reflected by the presence of noise, on token-level and sentence-level distillation. Using IWSLT14 de→en, IWSLT13 en→fr, WMT14 en→de, and IWSLT17 ar→en datasets, we aim to understand how various levels of noise influence the effectiveness of each distillation approach.

Experimental Setup and Methodology

To assess the impact of text complexity on knowledge distillation, we introduce varying levels of noise to the datasets. We follow the methodology in [Edunov *et al.*, 2018], applying three conditions to each dataset: *no* noise, *moderate* noise, and *high* noise. We introduce the noise through token manipulation including deletion, substitution, and swapping.

Specifically, under *moderate* noise conditions, we randomly delete and substitute 10% of the tokens and conduct token swapping with a 50% probability, maintaining a swap length of 3. This setup aims to simulate real-world linguistic processing errors and syntactic disarray. For *high* noise conditions, we keep the token deletion and substitution probabilities unchanged but increase the token swapping probability to 100%, further elevating syntactic complexity. Our implementation

[2]https://github.com/facebookresearch/fairseq

Dataset	Stud Size	BLEU Score					
		BS-Token	BS-Sentence	△BS	TF-Token	TF-Sentence	△TF
IWSLT14 de→en	3M	30.50	31.09	0.59	34.16	33.50	-0.66
IWSLT13 en→fr	12M	42.42	43.48	1.06	45.97	45.29	-0.68
WMT14 en→de	83M	26.49	26.77	0.28	29.82	28.58	-1.24
IWSLT17 ar→en	47M	32.18	31.15	1.03	32.51	31.60	-0.91

Table 3: Impact of decoding difficulty on BLEU scores: comparing Beam Search (BS) and Teacher Forcing (TF) methods. 'BS-Token' and 'BS-Sentence' represent BLEU scores using beam search for token-level and sentence-level distillation, respectively. 'TF-Token' and 'TF-Sentence' denote BLEU scores using teacher forcing for token-level and sentence-level distillation. △BS and △TF represent the differences in BLEU scores between token-level and sentence-level distillation for beam search and teacher forcing methods, respectively.

of these manipulations references the methods available in this resource[3]. In our experiments, the teacher models are Transformer-based, consistent with those in Table 1, using the default sizes in Fairseq [Ott *et al.*, 2019] for each dataset. Our analysis focuses on comparing results under different noise conditions to evaluate the impact of text complexity on distillation effectiveness. The results are shown in Table 2.

Analysis of Results and Summary

From the results in Table 2, we observe a trend across all datasets: as the text complexity increases, both token-level and sentence-level distillation show a decrease in performance. However, sentence-level distillation demonstrates greater resilience, evidenced by a generally smaller decline in BLEU scores compared to token-level distillation, particularly in high noise scenarios. This is reflected from the lower average △ Rate (S) across different noise levels, indicating its suitability for handling complex text scenarios. In contrast, token-level distillation exhibits a more significant performance drop with the increased text complexity, as shown by the higher △ Rate (T).

In general, when the noise is low, the token-level distillation shows higher accuracy than the sentence-level distillation (negative △ values in *Orig* noise setting in Table 2). As the noise become higher, student models trained with sentence-level distillation display a better performance than those with token-level distillation (positive △ values in *High* noise setting in Table 2). The above phenomenon aligns with our hypothesis that token-level distillation is more effective in simpler scenarios with lower text complexity.

These findings highlight the importance of text complexity in the selection of appropriate knowledge distillation methods for NMT. Sentence-level distillation emerges as a robust choice for complex text scenarios, while token-level distillation is preferable in simpler, less complex environments.

3.4 Impact of Decoding Difficulty

In this subsection, we examine the relationship between decoding difficulty and the performance of knowledge distillation methods. For decoding methods, we mainly take teacher forcing [Toomarian and Barhen, 1992; Lamb *et al.*, 2016] and beam search [Jaszkiewicz and Słowiński, 1999] into consideration. Beam search explores multiple hypotheses at each decoding step conditioned on the previous decoding results. Teacher forcing, different with beam search, directly uses the previous target sequence as condition at each step of sequence generation, effectively preventing error amplification during decoding. This method simplifies the decoding process and can lead to improved performance [Baskar *et al.*, 2019], which can be regarded as a simpler scenario in terms of decoding methods compared with the beam search.

Experimental Setup and Methodology

Experiments are conducted using the same datasets and teacher models as in Tables 1 and 2. The focus of our experiments is to closely examine the performance of token-level distillation and sentence-level distillation under different decoding difficulties (i.e., teacher forcing and beam search methods) on each dataset. Specifically, during the prediction phase, we employ beam search (BS) and teacher forcing (TF) methods. The former method considers the most probable candidates at each step of word generation, selecting one to include in the final sentence output. The latter method inputs the actual previous word into the model, rather than the model's own prediction from the previous step.

Analysis of Results and Summary

Table 3 presents a comparison of BLEU scores for both BS and TF methods across token-level and sentence-level distillation. Our results indicate that teacher forcing is more effective at the token-level compared to the sentence-level, as evidenced by the negative values in △TF across all datasets. This suggests that token-level distillation is better suited for the teacher forcing decoding approach.

Conversely, in the more complex beam search scenario, sentence-level distillation tends to outperform token-level distillation, as indicated by the positive values in △BS. This shift in effectiveness from token-level in TF to sentence-level in BS aligns with our hypothesis that teacher forcing, being a simpler decoding method, is more effective in scenarios where the decoding process is less complex. The token-level distillation benefits from the simplicity of the teacher forcing method, as it allows seeing the correct prefix words during decoding, making the process simpler and thus more effective.

[3]https://github.com/valentinmace/noisy-text/tree/e73c83dd1f08c25210c27abebf74d304de0d24e2

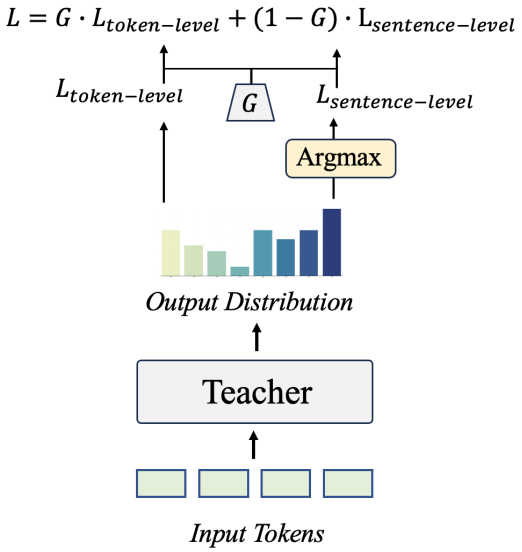

Figure 1: Architecture of the hybrid distillation method.

3.5 Summary

Based on our three comprehensive analyses focusing on model size, text complexity, and decoding difficulty, we have observed that token-level distillation is generally more suitable for scenarios involving larger student models, simpler texts, and greater amounts of available decoding information. In contrast, sentence-level distillation tends to be more effective in scenarios with smaller student models, more complex texts, and limited decoding information. These findings align with our initial hypothesis, suggesting that token-level distillation is better suited for simpler scenarios, while sentence-level distillation is more adept at handling complex situations.

4 Hybrid Method for Combining Token-Level and Sentence-Level Distillation

Despite our experimental results validate our hypothesis regarding the effectiveness of token-level and sentence-level distillation in different scenarios, we face the challenge of accurately defining the complexity level of each scenario. This issue complicates the optimal application of distillation methods in neural machine translation (NMT). In response, we propose a hybrid method, which combines token-level and sentence-level distillation through a dynamic gating mechanism. This method is designed to utilize the strengths of both distillation strategies and be adaptable across various scenarios, ranging from "simple" to "complex".

4.1 Hybrid Distillation Method

Our hybrid method features a gate-controlled mechanism, dynamically balancing the contributions of token-level and sentence-level distillation. This mechanism, denoted as G and illustrated in Figure 1, is represented by the function $g(x)$ for each input sequence x, modulating the influence of each distillation strategy during training to suit different translation scenarios.

The overall loss function, L, is a hybrid of token-level and sentence-level distillation losses, modulated by G. Let $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_m\}$ respectively represent the input and output (target) sequences. The probabilities $\mathrm{P_s}(y_j \mid x)$ and $\mathrm{P_t}(y_j \mid x)$ represent the output probabilities at position j for the student and teacher models, respectively.

For each input sequence x, the gate-controlled parameter $g(x)$ is defined as:

$$g(x) = \frac{1}{1 + e^{-z(x)}} \tag{1}$$

where $z(x)$ is a function of the input sequence x, determining the balance between token-level and sentence-level distillation for that particular input.

The token-level loss $L_{\text{token-level}}(x)$ is defined as:

$$L_{\text{token-level}}(x) = -\sum_{j=1}^{m} \sum_{y_j \in V} \mathrm{P_t}\left(y_j \mid x\right) \log \mathrm{P_s}\left(y_j \mid x\right) \tag{2}$$

which sums over all tokens y_j in the vocabulary V, weighted by the probability of teacher model $\mathrm{P_t}(y_j \mid x)$ and the logarithm of the probability of student model $\mathrm{P_s}(y_j \mid x)$.

The sentence-level loss $L_{\text{sentence-level}}(x)$ is defined as:

$$L_{\text{sentence-level}}(x) = -\log \mathrm{P_s}(\hat{y} \mid x) \tag{3}$$

which is the negative logarithm of the student model's probability of the actual output sequence $\hat{y}$ given by the teacher model.

Therefore, the overall loss function L for an input sequence x is given by:

$$L(x) = g(x) \cdot L_{\text{token-level}}(x) + (1 - g(x)) \cdot L_{\text{sentence-level}}(x) \tag{4}$$

This formulation allows $L(x)$ to represent the combined loss for a given input sequence x, effectively integrating the token-level and sentence-level distillation losses. By dynamically adjusting the weights of token-level and sentence-level distillation through $g(x)$, our hybrid method adapts to different input sequences, enhancing the effectiveness of model training.

4.2 Implementation Details

The training process begins with training a BiBERT teacher model at its base size to generate reference outputs. Subsequently, we implement our hybrid distillation method. This approach allows the model to adaptively switch between token-level and sentence-level strategies, optimizing the most effective learning path throughout the training process. Our experiments are conducted on four NVIDIA 3090 GPUs, each with a batch size of 3000. Gradients accumulate over four iterations per update. The learning rate is set at 5×10^{-4}, using the Adam optimizer with an inverse-sqrt learning rate scheduler. For inference, we employ a beam search with a width of 4 and a length penalty of 0.6.

4.3 Baselines

In our study, we compared our hybrid distillation approach with several advanced baseline methods in NMT:

- **Transformer + R-Drop** [Wu *et al.*, 2021]: Utilizes regularization to minimize the bidirectional KL-divergence between sub-models' outputs.

Methods	BLEU
Transformer + R-Drop [Wu *et al.*, 2021]	37.25
CipherDAug [Kambhatla *et al.*, 2022]	37.53
Cutoff [Shen *et al.*, 2020]	37.60
Cutoff+Knee [Iyer *et al.*, 2023]	37.78
SimCut [Gao *et al.*, 2022]	37.81
Transformer + R-Drop + Cutoff [Wu *et al.*, 2021]	37.90
Cutoff + R-A + LM [Lohrenz *et al.*, 2023]	37.96
Bi-SimCut [Gao *et al.*, 2022]	38.37
BiBERT [Xu *et al.*, 2021a]	38.61
Our Hybrid Distillation	**39.30**

Table 4: Experimental results on IWSLT14 de→en of baseline methods and our hybrid method.

- **CipherDAug** [Kambhatla *et al.*, 2022]: Employs a novel data augmentation technique based on ROT-k ciphers.
- **Cutoff** [Shen *et al.*, 2020]: Implements a data augmentation strategy that erases part of the information within an input sentence.
- **Cutoff+Knee** [Iyer *et al.*, 2023]: Combines Cutoff with an Explore-Exploit learning rate schedule.
- **SimCut and Bi-SimCut** [Gao *et al.*, 2022]: Enforces consistency between the output distributions of original and cutoff sentence pairs.
- **Transformer + R-Drop + Cutoff** [Wu *et al.*, 2021]: Integrates R-Drop regularization with Cutoff data augmentation.
- **Cutoff + Relaxed Attention + LM** [Lohrenz *et al.*, 2023]: Introduces relaxed attention as a regularization technique.
- **BiBERT** [Xu *et al.*, 2021a]: Utilizes a bilingual pretrained language model for the NMT encoder.

4.4 Experimental Results

Table 4 shows the translation accuracy (indicated by BLEU score) of our method and baseline methods. The results demonstrate that our hybrid distillation method outperforms all baseline models, achieving a BLEU score of 39.30, which indicates the efficiency of our method in combining token-level and sentence-level distillation strategies.

4.5 Ablation Study

The ablation study evaluates the individual impacts of token-level and sentence-level distillation within our hybrid method, aiming to understand their contributions to the overall translation performance.

Table 5 presents the results of the ablation study. The individual performances of sentence-level and token-level distillation highlight their respective strengths in enhancing translation quality. The sentence-level method, with a BLEU score of 39.01, demonstrates its capability in capturing the overall semantic coherence, while the token-level method, scoring slightly higher at 39.15, shows its effectiveness in ensuring precise token-level translations. Our hybrid method, achieving a BLEU score of 39.30, surpasses these individual strategies, indicating that the synergistic combination of token-level precision and sentence-level coherence can yield superior results. The results show our hybrid method, which combines token-level and sentence-level distillation, effectively navigates the challenges in scenarios with ambiguous complexity levels, enhancing translation quality in neural machine translation.

Methods	Model Params	BLEU
Sentence-Level	78M	39.01
Token-level	78M	39.15
Our Hybrid Distillation	78M	39.30

Table 5: Ablation study results of distillation methods on IWSLT14 de→en.

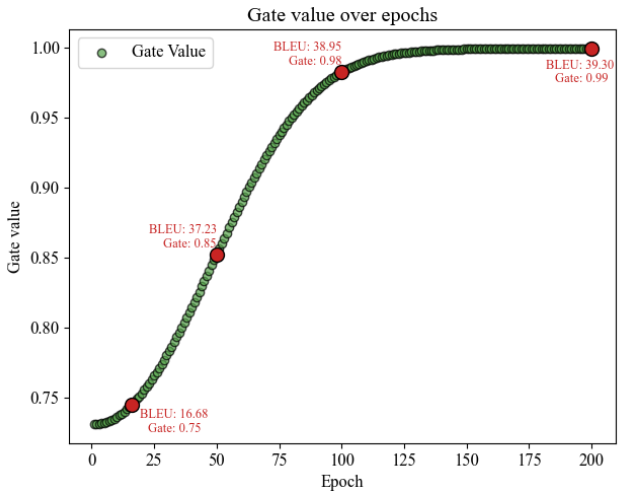

Figure 2: Dynamics of gate value G over training epochs.

4.6 Analysis of Gate-Controlled Mechanism

To understand the learning process of the learnable gate-controlled mechanism G and to verify the effectiveness of this learning method, we present the dynamics of the gate value G over training epochs during our experiments, as shown in Figure 2. We find that at the beginning of the learning process of G, its value is around 0.72. As training progresses (around 20 epochs), the value of G increases to 0.75, with the corresponding BLEU score being 16.68. With further training (around 50 epochs), G gradually rises to 0.85, and the BLEU score significantly improves to 37.23. During this phase, the increase in the value of G is quite apparent, and there is a notable enhancement in the BLEU score. Subsequently (around 100 epochs), G increases to 0.98, and the BLEU score rises to 38.95. At this stage, although G continues to increase, the growth rate of the BLEU score slows down compared to the previous phase. Eventually, the value of G approaches 1, and the BLEU score reaches 39.30. We believe that initially, sentence-level learning is easier, while token-level learning is more challenging. Therefore, the model first learns the simpler aspects, leading to a faster increase in the BLEU score. As the simpler tasks are mastered, the model then moves on to

the more difficult token-level learning, resulting in a slower rate of improvement in the BLEU score. From the results, it is evident that the learnable parameters, by adjusting the size of G, effectively enable the model to autonomously learn knowledge from sentence-level distillation and token-level distillation, demonstrating the effectiveness of our design.

5 Conclusion

In this paper, we conduct an in-depth exploration of the two main methods of knowledge distillation in neural machine translation (NMT): sentence-level and token-level distillation. We hypothesize that token-level distillation is more suitable for simpler scenarios, whereas sentence-level distillation is better for complex scenarios. To test this hypothesis, we systematically analyze the impact of varying the size of the student model, the complexity of the text, and the difficulty of the decoding process. Our empirical results validate our hypothesis, showing that token-level distillation generally performs better in scenarios with larger student models, simpler texts, and higher availability of decoding information (making decoding easier). In contrast, sentence-level distillation performs better in scenarios with smaller student models, more complex texts, and limited decoding information (making decoding harder). To address the challenge of defining the difficulty level of specific scenarios, we further introduce a dynamic gate-controlled mechanism that combines the advantages of both token-level and sentence-level distillation. Our experiments validate the effectiveness of this hybrid method over the single distillation method and baselines methods.

Ethical Statement

There are no ethical issues.

Acknowledgments

We are grateful to the anonymous reviewers of IJCAI for their constructive comments that significantly improve the manuscript. This work is supported by the Liaoning Provincial Applied Basic Research Program, grant number 2022JH2/101300258.

References

[Baskar *et al.*, 2019] Murali Karthick Baskar, Lukáš Burget, Shinji Watanabe, Martin Karafiát, Takaaki Hori, and Jan Honza Černockỳ. Promising accurate prefix boosting for sequence-to-sequence asr. In *ICASSP*, pages 5646–5650. IEEE, 2019.

[Blain *et al.*, 2023] Frédéric Blain, Chrysoula Zerva, Ricardo Ribeiro, Nuno M Guerreiro, Diptesh Kanojia, José GC de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, et al. Findings of the wmt 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, 2023.

[Chen *et al.*, 2020] Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. Distilling knowledge learned in bert for text generation. In *ACL*, pages 7893–7905, 2020.

[Deng *et al.*, 2023] Hexuan Deng, Liang Ding, Xuebo Liu, Meishan Zhang, Dacheng Tao, and Min Zhang. Improving simultaneous machine translation with monolingual data. In *AAAI*, pages 12728–12736, 2023.

[Ding *et al.*, 2020] Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. Understanding and improving lexical choice in non-autoregressive translation. In *ICLR*, 2020.

[Ding *et al.*, 2021] Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. Progressive multi-granularity training for non-autoregressive translation. In *Findings of the ACL-IJCNLP 2021*, pages 2797–2803, 2021.

[Edunov *et al.*, 2018] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *EMNLP*, pages 489–500, 2018.

[Farinha *et al.*, 2022] Ana C Farinha, M Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José GC De Souza, Helena Moniz, and André FT Martins. Findings of the wmt 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, 2022.

[Freitag *et al.*, 2017] Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*, 2017.

[Gajbhiye *et al.*, 2021] Amit Gajbhiye, Marina Fomicheva, Fernando Alva-Manchego, Frédéric Blain, Abiola Obamuyide, Nikolaos Aletras, and Lucia Specia. Knowledge distillation for quality estimation. In *ACL*, pages 5091–5099, 2021.

[Gao *et al.*, 2022] Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. Bi-simcut: A simple strategy for boosting neural machine translation. In *NAACL*, pages 3938–3948, 2022.

[Ge *et al.*, 2023] Ling Ge, Chunming Hu, Guanghui Ma, Hong Zhang, and Jihong Liu. Prokd: an unsupervised prototypical knowledge distillation network for zero-resource cross-lingual named entity recognition. In *AAAI*, pages 12818–12826, 2023.

[Gou *et al.*, 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[He *et al.*, 2021] Haoyu He, Xingjian Shi, Jonas Mueller, Sheng Zha, Mu Li, and George Karypis. Distiller: A systematic study of model distillation methods in natural language processing. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 119–133, 2021.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Iyer *et al.*, 2023] Nikhil Iyer, V. Thejas, Nipun Kwatra, Ramachandran Ramjee, and Muthian Sivathanu. Wide-minima density hypothesis and the explore-exploit learning

rate schedule. *Journal of Machine Learning Research*, 24(65):1–37, 2023.

[Jaszkiewicz and Słowiński, 1999] Andrzej Jaszkiewicz and Roman Słowiński. The 'light beam search'approach–an overview of methodology applications. *European Journal of Operational Research*, 113(2):300–314, 1999.

[Kambhatla *et al.*, 2022] Nishant Kambhatla, Logan Born, and Anoop Sarkar. Cipherdaug: Ciphertext based data augmentation for neural machine translation. In *Proceedings of the ACL*, pages 201–218, 2022.

[Kim and Rush, 2016] Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *EMNLP*, pages 1317–1327, 2016.

[Lamb *et al.*, 2016] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *NeurIPS*, 29, 2016.

[Lee *et al.*, 2022] Dongkyu Lee, Zhiliang Tian, Yingxiu Zhao, Ka Chun Cheung, and Nevin Zhang. Hard gate knowledge distillation-leverage calibration for robust and reliable language model. In *EMNLP*, pages 9793–9803, 2022.

[Lei *et al.*, 2022] Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. Sentence-level media bias analysis informed by discourse structures. In *EMNLP*, pages 10040–10050, 2022.

[Li *et al.*, 2021] Zheng Li, Danqing Zhang, Tianyu Cao, Ying Wei, Yiwei Song, and Bing Yin. Metats: Meta teacher-student network for multilingual sequence labeling with minimal supervision. In *EMNLP*, pages 3183–3196, 2021.

[Liao *et al.*, 2020] Baohao Liao, Yingbo Gao, and Hermann Ney. Multi-agent mutual learning at sentence-level and token-level for neural machine translation. In *EMNLP*, pages 1715–1724, 2020.

[Liu *et al.*, 2020] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020.

[Lohrenz *et al.*, 2023] Timo Lohrenz, Björn Möller, Zhengyang Li, and Tim Fingscheidt. Relaxed attention for transformer models. In *IJCNN*, pages 1–10, 2023.

[Ma *et al.*, 2023] Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. Multi-teacher knowledge distillation for end-to-end text image machine translation. In *ICDAR*, pages 484–501. Springer, 2023.

[Mhamdi *et al.*, 2023] Meryem Mhamdi, Jonathan May, Franck Dernoncourt, Trung Bui, and Seunghyun Yoon. Multilingual sentence-level semantic search using meta-distillation learning. *arXiv preprint arXiv:2309.08185*, 2023.

[Mun'im *et al.*, 2019] Raden Mu'az Mun'im, Nakamasa Inoue, and Koichi Shinoda. Sequence-level knowledge distillation for model compression of attention-based sequence-to-sequence speech recognition. In *ICASSP*, pages 6151–6155, 2019.

[Oka *et al.*, 2021] Yui Oka, Katsuhito Sudoh, and Satoshi Nakamura. Using perturbed length-aware positional encoding for non-autoregressive neural machine translation. *arXiv preprint arXiv:2107.13689*, 2021.

[Ott *et al.*, 2019] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*, pages 48–53, 2019.

[Peng *et al.*, 2023] Keqin Peng, Liang Ding, Qihuang Zhong, Yuanxin Ouyang, Wenge Rong, Zhang Xiong, and Dacheng Tao. Token-level self-evolution training for sequence-to-sequence learning. In *Proceedings of the ACL*, pages 841–850, 2023.

[Phuong and Lampert, 2019] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *ICML*, pages 5142–5151. PMLR, 2019.

[Ren *et al.*, 2019] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *NeurIPS*, 32, 2019.

[Shao *et al.*, 2022] Chenze Shao, Xuanfu Wu, and Yang Feng. One reference is not enough: Diverse distillation with reference selection for non-autoregressive translation. In *NAACL*, pages 3779–3791, 2022.

[Shen *et al.*, 2020] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*, 2020.

[Stahlberg, 2020] Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.

[Sun *et al.*, 2019] Hao Sun, Xu Tan, Jun-Wei Gan, Hongzhi Liu, Sheng Zhao, Tao Qin, and Tie-Yan Liu. Token-level ensemble distillation for grapheme-to-phoneme conversion. In *ISCA*, pages 2115–2119, 2019.

[Sun *et al.*, 2020] Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. Knowledge distillation for multilingual unsupervised neural machine translation. In *ACL*, pages 3525–3535, 2020.

[Tan *et al.*, 2019] Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. In *ICLR*, 2019.

[Tan *et al.*, 2022] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *ACL*, pages 1672–1681, 2022.

[Tang *et al.*, 2019] Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*, 2019.

[Tang *et al.*, 2021] Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *NeurIPS*, 34:24468–24481, 2021.

[Toomarian and Barhen, 1992] Nikzad Benny Toomarian and Jacob Barhen. Learning a trajectory using adjoint functions and teacher forcing. *Neural networks*, 5(3):473–484, 1992.

[Wang and Yoon, 2021] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068, 2021.

[Wang *et al.*, 2020] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu. Structure-level knowledge distillation for multilingual sequence labeling. In *ACL*, pages 3317–3330, 2020.

[Wang *et al.*, 2021] Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. Selective knowledge distillation for neural machine translation. In *ACL and IJCAI*, pages 6456–6466, 2021.

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

[Xiao *et al.*, 2023] Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-yan Liu. A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Xu *et al.*, 2021a] Haoran Xu, Benjamin Van Durme, and Kenton W. Murray. Bert, mbert, or bibert? A study on contextualized embeddings for neural machine translation. In *EMNLP*, pages 6663–6675, 2021.

[Xu *et al.*, 2021b] Weijia Xu, Shuming Ma, Dongdong Zhang, and Marine Carpuat. How does distilled data complexity impact the quality and confidence of non-autoregressive machine translation? In *ACL*, pages 4392–4400, 2021.

[Yang *et al.*, 2022a] Zhixian Yang, Renliang Sun, and Xiaojun Wan. Nearest neighbor knowledge distillation for neural machine translation. In *NAACL*, pages 5546–5556, 2022.

[Yang *et al.*, 2022b] Zhixian Yang, Renliang Sun, and Xiaojun Wan. Nearest neighbor knowledge distillation for neural machine translation. In *NAACL*, pages 5546–5556, 2022.

[Zhang *et al.*, 2019] Biao Zhang, Deyi Xiong, Jinsong Su, and Jiebo Luo. Future-aware knowledge distillation for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2278–2287, 2019.

[Zhang *et al.*, 2022] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *CVPR*, pages 10174–10183, 2022.

[Zhang *et al.*, 2024] Shuoxi Zhang, Hanpeng Liu, and Kun He. Knowledge distillation via token-level relationship graph based on the big data technologies. *Big Data Research*, page 100438, 2024.

[Zhou *et al.*, 2020] Chunting Zhou, Jiatao Gu, and Graham Neubig. Understanding knowledge distillation in non-autoregressive machine translation. In *ICLR*, 2020.