TAeKD: Teacher Assistant Enhanced Knowledge Distillation for Closed-Source Multilingual Neural Machine Translation

Bo Lv[1,2,3], Kaiwen Wei[3,4], Xin Liu[2*], Ping Luo[1,2,3*], Yue Yu[2*]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]Peng Cheng Laboratory, Shenzhen 518066, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]College of Computer Science, Chongqing University, Chongqing, 400044, China
{lvbo19,weikaiwen19}@mails.ucas.ac.cn
hit.liuxin@gmail.com, luop@ict.ac.cn, yuy@pcl.ac.cn

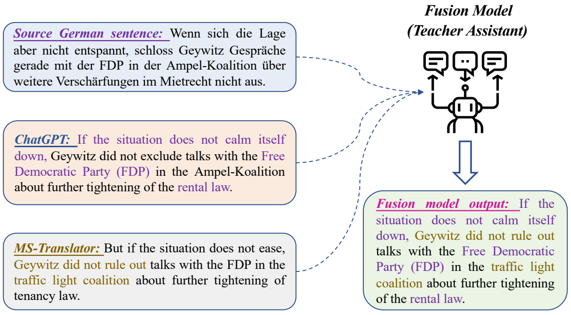
Abstract

Knowledge Distillation (KD) serves as an efficient method for transferring language knowledge from open-source large language models (LLMs) to more computationally efficient models. However, challenges arise when attempting to apply vanilla KD methods to transfer knowledge from closed-source Multilingual Neural Machine Translation (MNMT) models based on LLMs. In this scenario, the soft labels and training data are not accessible, making it difficult to achieve effective knowledge transfer. To address this issue, this paper proposes a Teacher Assistant enhanced Knowledge Distillation (TAeKD) method to augment the knowledge transfer capacity from closed-source MNMT models. Specifically, TAeKD designs a fusion model that integrates translation outputs from multiple closed-source models to generate soft labels and training samples. Furthermore, a quality assessment learning mechanism is introduced to enhance the generalization of the fusion model and elevate the quality of the fusion data used to train the student model. To facilitate research on knowledge transfer from MNMT models, we also introduce FuseData, a benchmark consisting of a blend of translations from multiple closed-source systems. The experimental results show that TAeKD outperforms the previous state-of-the-art KD methods on both WMT22 and FLORES-101 test sets.

Keywords: Knowledge distillation, Closed-source, Multilingual neural machine translation

1. Introduction

Large language models (LLMs) (Xue et al., 2021; Brown et al., 2020) have achieved increasingly impressive results in the field of Multilingual Neural Machine Translation (MNMT) (Dabre et al., 2021). This has inspired researchers to focus on transferring LLM knowledge to smaller and more computationally efficient models. Knowledge distillation (KD) (Hinton et al., 2015; Gou et al., 2021) is widely regarded as an effective method for transferring knowledge from large models to smaller networks. The vanilla KD technique involves training a student model by utilizing two types of labels: soft labels from the teacher model (i.e., the probabilities of candidate tokens with a temperature coefficient) and the correct training data labels (i.e., hard labels). However, due to commercial reasons, the state-of-the-art LLM-based MNMT systems that support hundreds of language pairs, such as ChatGPT[1] (OpenAI, 2022) and Microsoft (MS) Translator[2], are typically closed-source, which means that their soft labels and training data are not accessible to the public. This dilemma hinders researchers from employing the vanilla KD approach for transferring knowledge from these closed-source MNMT systems.



Source German sentence: Wenn sich die Lage aber nicht entspannt, schloss Geywitz Gespräche gerade mit der FDP in der Ampel-Koalition über weitere Verschärfungen im Mietrecht nicht aus.

ChatGPT: If the situation does not calm itself down, Geywitz did not exclude talks with the Free Democratic Party (FDP) in the Ampel-Koalition about further tightening of the rental law.

MS-Translator: But if the situation does not ease, Geywitz did not rule out talks with the FDP in the traffic light coalition about further tightening of tenancy law.

Fusion model output: If the situation does not calm itself down, Geywitz did not rule out talks with the Free Democratic Party (FDP) in the traffic light coalition about further tightening of the rental law.

Figure 1: An example of the fusion model as a teacher assistant to fine-grained fuse the ChatGPT translation and MS-Translator translation.

Recent studies predominantly employ sequence-level KD (Kim and Rush, 2016) with multiple teachers to address this problem. The sequence-level KD can be considered a form of data augmentation, similar to back-translation (Sennrich et al., 2016). In this method, multiple closed-source MNMT systems are used to generate candidate translations, utilizing the same monolingual data as input. Subsequently, a student model is trained on the translation with the highest score among the candidate translations. The score

* Corresponding author.

[1] https://chat.openai.com/

[2] http://www.microsoft.com/translator

LREC-COLING 2024, pages 15530–15541
20-25 May, 2024.

can be calculated using a multilingual sentence embedding model (Feng et al., 2022) that utilizes the source data as a reference.

However, existing sequence-level KD methods have two limitations: (1) Due to the differences in parameters and architectures, MNMT systems exhibit diverse strengths and weaknesses (Guerreiro et al., 2023), making it difficult to effectively blend their respective strengths through sentence-level selection. For example, ChatGPT translations exhibit fewer grammar errors and higher fluency, while Microsoft translations adhere more closely to the source text. (2) Previous efforts (Tan et al., 2019) have demonstrated that soft labels are the primary reason for the high knowledge transfer efficacy in vanilla knowledge distillation methods (Hinton et al., 2015). The soft labels achieve a smoother distribution and enhance the learning process of the student model by transmitting similarity information between tokens with nonzero probability. Therefore, the absence of soft labels results in the inadequate transfer of knowledge from the closed-source MNMT systems to the student model.

To alleviate the limitations mentioned above, in this work, we introduce a **T**eacher **A**ssistant **e**nhanced **K**nowledge **D**istillation (TAeKD) method that transfers more knowledge from closed-source MNMT systems to the student. Specifically, TAeKD consists of two parts: (a) the fusion model, and (b) knowledge distillation using the fusion model as a teacher assistant. In the first part, TAeKD designs a fusion model that can fine-grained integrate the candidate translations from different MNMT systems. Since the parameters of the fusion model are accessible, it can provide soft labels that convey similarity information between tokens of different candidate translations. In addition, to enhance the fusion model's capacity to assess the quality of candidate translations, we introduce a Quality Assessment Learning mechanism (QAL). This mechanism empowers the fusion model to generate higher-quality fusion data. As shown in Figure 1, the fusion model takes the source sentence, along with the candidate translations generated by ChatGPT and MS-translator as inputs, and then produces a finely-fused translation. In the second part, TAeKD utilizes the fusion model as a teacher assistant (Mirzadeh et al., 2020) to generate high-quality fine-grained fusion data and corresponding soft labels for training student model. By this way, TAeKD can efficiently transfer knowledge to the student model at both the sequence-level and word-level.

To assess the effectiveness of the knowledge distillation approach for closed-source models, we introduce a dataset named FuseData. This dataset comprises six different translation directions, with each direction containing 100k samples for training the fusion model and over 150k samples for training the student model. Each source text undergoes translation using the ChatGPT and MS-Translator MNMT systems, and the training data for the fusion model also includes ground-truth labels.

We evaluate the proposed TAeKD method on WMT22 (Kocmi et al., 2022) and FLORES-101 (Goyal et al., 2022) test sets at six different translation directions. Experimental results show that TAeKD significantly outperforms previous KD methods in both BLEU and COMET-22 metrics. In summary, the contributions of the paper are as follows:

- This paper delves into the knowledge transfer process from closed-source MNMT systems. We introduce a Teacher Assistant enhanced Knowledge Distillation method (TAeKD), which can finely fuse candidate translations and supply soft labels to enhance distillation efficacy.
- This paper proposes a Quality Assessment Learning (QAL) mechanism to enhance the fusion model's generalization and improve the quality of fusion data for training the student model.
- This paper introduces the FuseData dataset, comprising six translation directions with over 2 million samples. Experiments on both WMT22 and FLORES-101 test sets demonstrate a stronger generalization ability of the proposed TAeKD method. The code and dataset are publicly available for research purposes[3].

2. Preliminaries

2.1. Multilingual Neural Machine Translation

MNMT models are capable of translating between multiple language pairs (Dabre et al., 2021). Given a source sentence with N tokens $s = \{x_1, x_2, \ldots x_N\}$ and the corresponding target sentence with M tokens $t = \{y_1, y_2, \ldots y_M\}$, the training objective for MNMT models is maximize the probability of each target token conditioning on the source sentence by the cross-entropy (CE) loss:

$$\mathcal{L}_{ce} = -\sum_{j=1}^{M} \log p(y_j | y_{<j}, x; \theta) \tag{1}$$

where y_j denotes the ground-truth target, $y_{<j}$ denotes the target-side previous context at time step j, and θ denotes the model parameters.

2.2. Word-Level Knowledge Distillation

In vanilla knowledge distillation (Hinton et al., 2015), also known as word-level distillation (Kim and Rush,

[3] https://github.com/lvbotenbest/TAeKD

2016), the student model matches both the outputs of the ground-truth one-hot label and the soft labels provided by the teacher model. The cross entropy between two distributions serves as the distillation loss:

$$\mathcal{L}_{word-kd} = \sum_{j}^{M} \sum_{k=1}^{|\nu|} q\left\{y_j^* = k|y_{<j}, x, \theta_T\right\} \times \log p(y_j^* = k|y_{<j}, x, \theta_S) \tag{2}$$

where the θ_T and θ_S denote the model parameters of the teacher and the student, respectively. The $|\nu|$ denotes the vocabulary size of the target language. The $q\left\{y_j^* = k|y_{<j}, x, \theta_T\right\}$ is the soft label of the teacher model for token k at j-th step, and is calculated by softening the model's output distribution with temperature τ as follow:

$$q\left\{y_j^* = k|y_{<j}, x, \theta_T\right\} = \frac{exp(z_k/\tau)}{\sum_i exp(z_i/\tau)} \tag{3}$$

where the z_k and z_i are the probabilities predicted by the teacher model for candidate tokens, which correspond to the k-th and i-th tokens in the vocabulary, respectively.

Then, the overall loss function of word-level KD is as follows:

$$\mathcal{L}_{kd} = (1-\alpha)\mathcal{L}_{ce} + \alpha\mathcal{L}_{word-kd} \tag{4}$$

where α is the weight coefficient.

2.3. Sequence-level Knowledge Distillation

Sequence-level KD (Kim and Rush, 2016) encourages the student model to imitate the sequence probabilities of the translations from the teacher model. To this end, it optimizes the student model through the following CE loss:

$$\mathcal{L}_{seq-kd} = -\sum_{j=1}^{m} \log p(\widehat{y}_j|\widehat{y}_{<j}, x; \theta) \tag{5}$$

where m is the sequence length of the translation generated by the teacher, and $\widehat{y}_j$ denotes the generated target by the teacher at time step j.

2.4. Quality-Aware Sequence-level Knowledge Distillation

The training data generated by the translation system are not error-free, and such errors may have a significant impact on the training effectiveness of the student model. To this end, the Quality-Aware Sequence-level KD (team et al., 2022) utilizes N translation systems to translate the same monolingual text, denoted as src. Subsequently, a multilingual encoding model $E(\cdot)$, such as Labse (Feng et al., 2022), is utilized to calculate the similarity between the translations and the src, taking the highest-scoring translation as training data. The similarity score is defined as:

$$score_i = cos(E(src), E(y_i)) \tag{6}$$

where y_i is the translation of src translated by the i th MNMT system. $E(src)$ and $E(y_i)$ are the sentence embedding of src and y_i. For $N = 2$, the highest-scoring translation is selected by $argmax(score_1, score_2)$.

3. Methodology

In this section, we elaborate on our proposed Teacher Assistant enhanced Knowledge Distillation method (TAeKD). First, we introduce the fusion model and the Quality Assessment Learning (QAL) mechanism for enhancing the performance of the fusion model. Then, we explain the process of utilizing the fusion model as a teacher assistant to train a student model. The overall framework of TAeKD is illustrated in Figure 2.

3.1. Fusion Model

The objective is to devise an open-source generative model as a teaching assistant that takes the source sentence src along with its different translations $\{t_1, t_2\}$ as input, and produces an enhanced output $\hat{t}$ and the soft labels. To accomplish this, we present the fusion model, an encoder-decoder approach designed to combine candidate translations generated by ChatGPT and MS translator systems at a fine-grained level. Specifically, we concatenate the source sentence and the candidates translations using separator tokens, such as <source sentence is:>,<system A translation is:>. We fine-tune two mt0-large model (Muennighoff et al., 2023) for English→German/Russian/Czech and German/Russian/Czech→English. We add different prompts for different translation directions. For instance, considering the fusing from English→German, the input template is:

$$\begin{aligned}&Translate\ next\ English\ sentence\ to\ German:\\&source\ sentence\ is:\{\text{source sentence}\};\\&system\ A\ translation\ is:\{\text{translation A}\};\\&system\ B\ translation\ is:\{\text{translation B}\}\end{aligned}$$

We tried various different prompts and found that as long as the description of the language direction is added, other variations in the template have little impact on the model's performance.

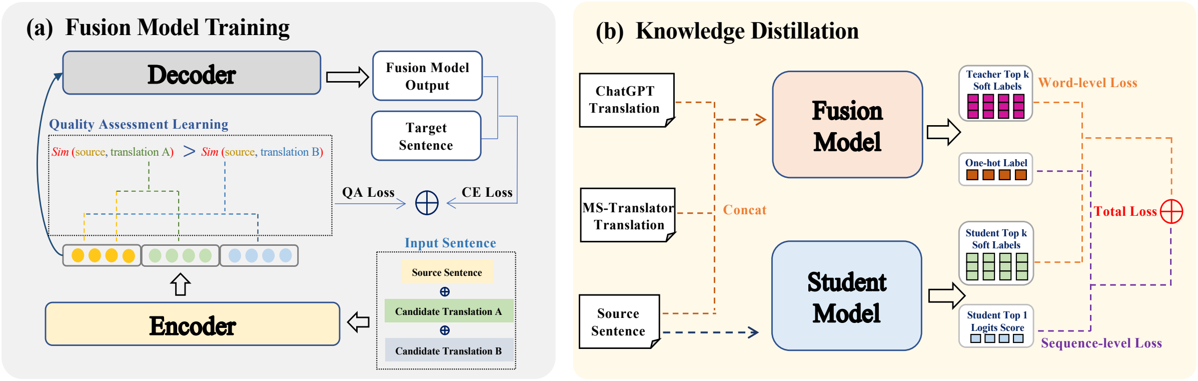


Figure 2: The overview of the proposed TAeKD method, which includes: (a) training the fusion model, and (b) conducting knowledge distillation using the trained fusion model as the teacher assistant. During the process of knowledge distillation, the source language is initially inputted into ChatGPT and MS Translator to generate translations. These translations are then combined and fed into the fusion model. Finally, the outputs of the fusion model are utilized to train the student model at both the word-level and sequence-level.

3.2. Quality Assessment Learning Mechanism

The main goal of the fusion model is to integrate the advantages of candidate translations. To achieve this goal, it is crucial for the fusion model to possess a strong ability to discriminate translation quality, enabling it to focus its attention on each candidate translation fragment with the highest quality during the decoding stage, thereby generating high-quality training data. Nonetheless, the highest quality candidate translation fragment may not necessarily appear in the training labels. Optimizing the model solely based on cross-entropy loss to fit the training labels can result in the model failing to learn to distinguish which candidate translation is better. To address this issue, we introduce the Quality Assessment learning during the training process, explicitly teaching the model to rank the translation quality of the input systems.

Before training, QAL uses Comet-Compare (Rei et al., 2022a) to rank the candidate translations and we obtain their ranking order. Comet-Compare is a translation quality evaluation tool that assesses the quality of translation results by evaluating the alignment between the translation result with the source text and reference translation at both the word and sentence levels. We add rank loss (Wang et al., 2019) to the encoder component of the fusion model, optimizing the embedding by considering the alignment degree between the candidate translations and the source text. By utilizing this training mechanism, the integrated model can produce high-quality training data during the decoding process by placing more attention on candidate translation fragments that display higher similarity with the source text embedding.

Formally, given a sequence pair (x, y) with two candidate translations C^1, C^2, where $x = (x_1, \ldots, x_N)$ is the source sentence of length N, $y = (y_1, \ldots, y_L)$ is the label sentence of length L and $C^1 = (C^1_1, \ldots, C^1_M)$ of length M, we convert the source sentence and candidate translations into the template in Section 3.1, and obtain the input $\mathcal{X}$.Then, we feed $\mathcal{X}$ into the fusion model $f_T = (f_T^{enc}, f_T^{dec})$. The encoder embeddings of x and C^1, C^2 are $H_x = \{h_{x_1}, \ldots, h_{x_S}\}$, $H_{C^1} = \{h_{C^1_1}, \ldots, h_{C^1_M}\}$ and $H_{C^2} = \{h_{C^2_1}, \ldots, h_{C^2_N}\}$. We take the average of all the word vectors in the source text and candidate translations, and use cosine to calculate the similarity between them:

$$sim(x, C^i) = cos(\frac{H_x}{\mathcal{S}}, \frac{H_{C^i}}{\mathcal{Q}}) \tag{7}$$

where $\mathcal{S}$ and $\mathcal{Q}$ are the token length of H_x and H_{C^i}. To assign higher similarity scores to better candidate translations and smaller scores to worse ones, we employ the rank loss for optimization:

$$\mathcal{L}_{qa} = \begin{cases} max(0, sim(x, C^1) - sim(x, C^2)) & r_1 < r_2 \\ max(0, sim(x, C^2) - sim(x, C^1)) & r_2 < r_1 \end{cases} \tag{8}$$

Then, we add this loss to the original cross-entropy loss :

$$\mathcal{L}_{ce} = -\sum_{j=1}^{L} \log p(y_j | y_{<j}, \mathcal{X}; \theta_T) \tag{9}$$

where θ_T is the parameter of the fusion model. The total loss becomes:

$$\mathcal{L}_{all} = \mathcal{L}_{ce} + \beta \mathcal{L}_{qa} \tag{10}$$

where β is the weight coefficient.

3.3. Fusion Model as a Teacher Assistant for Knowledge Distillation

Before training the student model, we utilize the already trained fusion model to generate high-quality training samples based on the candidate translations generated by multiple close-source systems. In order to save GPU memory during the training of the student model, we save the top-K probabilities of each step in the fusion model generation process and normalize them, making their sum equal to 1 for distillation. This can reduce the memory cost from the scale of $|\nu|$ to K.

Let x denote the source text, $\mathcal{X}$ denote the input of fusion model obtained by inputting x and candidate translations into the template of Section 3.1, $\widehat{Y}$ denote the target sentence generated by the fusion model ($\widehat{Y} = f_T(\mathcal{X})$). Let q denote the top-K soft labels, which can be obtained as follows:

$$q\left\{Y_j^* = n|\widehat{Y}_{<j}, x, \theta_T\right\} = \frac{exp(z_n/\tau)}{\sum_K exp(z_i/\tau)} \quad (11)$$

where the z_n and z_i are the probabilities for the n-th and i-th token in the vocabulary, respectively. τ is a hyperparameter for softening the model's output top-K probabilities z. The word-level KD loss can be expressed as:

$$\begin{aligned}\mathcal{L}_{word-kd} = \sum_{j}^{M}\sum_{i=1}^{k} q\left\{Y_j^* = i|\widehat{Y}_{<j}, x, \theta_T\right\} \times \\ \log p(Y_j^* = i|\widehat{Y}_{<j}, x, \theta_S)\end{aligned} \quad (12)$$

where M is the length of $\widehat{Y}$. p is derived by utilizing the same softening approach on the student model's output probabilities as is used in q. Then, the overall loss function of KD is the linear interpolation between the sequence-level KD loss and the word-level KD loss:

$$\begin{aligned}\mathcal{L}_{seq-kd} &= -\sum_{j=1}^{m} \log p(\widehat{Y}_j|\widehat{Y}_{<j}, x; \theta) \\ \mathcal{L}_{kd} &= (1-\lambda)\mathcal{L}_{seq-kd} + \lambda\mathcal{L}_{word-kd}\end{aligned} \quad (13)$$

where λ is the coefficient to trade off the two loss terms.

4. Experiments

4.1. Experimental Settings

Datasets. To facilitate research on knowledge transfer from closed-source systems, this paper releases a new dataset, namely FuseData, which consists of a blend of translations from two closed-source systems, ChatGPT and MS translator. FuseData includes six language pairs: English (en)->German (de), English->Russian (ru), English ->Czech (cs), and their reverses. As shown in Table 1, the FuseData package comprises two parts of the training set: one part is used for training the fusion model, while the other is used for training the student model. The data used for training the fusion model comes from the European Parallel corpus[4] (en<->de, en<->cs) and the United Nations Parallel corpus[5] (en<->ru), because these two parallel corpora are of relatively high-quality. We then utilize the LaBSE (Feng et al., 2022) to filter out high-quality parallel corpora with scores greater than 0.88. Since the en<->cs direction resulted in only 100k data after screening, to balance the total amount of data in each language direction, we select the top 100k parallel corpora in terms of LaBSE score for en<->de, en<->ru. We then input the source language of each direction into ChatGPT and MS-Translator for translation to obtain the data for training the fusion model. The data for training the student model originates from the single-language data filtered from WMT16[6] (there are no repetitions with the data used for training the fusion model), which is obtained after the translation by ChatGPT and MS-Translator. When filtering this portion of data, we first use the Jaccard Distance (Hancock, 2004) to deduplicate texts with a similarity greater than 0.8, and then use GPT2 (Radford et al., 2019) to calculate the perplexity of sentences in the en->xx[7] direction, selecting those that have a perplexity of less than 200 (the lower the perplexity, the smoother the text). For the xx->en direction, we employ XLM-R (Conneau et al., 2020) to select in the same manner.

Part	en->de	en->cs	en->ru	sum
fusion	100k	100k	100k	300k
student	253.0k	231.3k	237.2k	721.5k
Part	**de->en**	**cs->en**	**ru->en**	**sum**
fusion	100k	100k	100k	300k
student	213.1k	126.3k	162.1k	501.5k

Table 1: The statistics of the FuseData dataset.

For the fusion model, we randomly select 2,000 instances per language direction from the fusion part of FuseData. These instances are designated as the validation and test sets.

For the student model, we select 1,000 instances randomly from each language direction in the student part and use these as the validation set. We use the public benchmarks from WMT22 (Kocmi et al., 2022) and FLORES-101 (Goyal et al., 2022)

[4] https://www.statmt.org/europarl

[5] https://conferences.unite.un.org/uncorpus

[6] https://huggingface.co/datasets/wmt16

[7] 'xx' refers to the German, Czech, and Russian languages as a whole.

Lang-Pair	WMT22 sentences	FLORES-101 sentences
cs->en	1448	1012
en->cs	2037	1012
de->en	1984	1012
en->de	2037	1012
ru->en	2016	1012
en->ru	2037	1012

Table 2: The number of sentences contained in the test datasets used for evaluation.

test datasets for evaluation. As shown in Table 2, these two test datasets have 1000 to 2000 test sentences in each language direction.

Baselines. The baselines we used for comparison are the Sequence-level Knowledge Distillation (SLKD) methods (Kim and Rush, 2016; Gordon and Duh, 2019) and the Quality-Aware Sequence-level Knowledge Distillation (QA-SLKD) methods. In this paper, **ChatGPT-SL** and **MS-SL** represent SLKD methods using the translation results from ChatGPT and MS-Translator, respectively. **LaBSE-QA** refers to the QA-SLKD method that uses LaBSE (Feng et al., 2022) as a quality assessment model to filter high-quality translations from candidate translations for training data. **Comet-QA**[8]**.** refers to the QA-SLKD method based on the cometkiwi-da (Rei et al., 2022b) quality assessment model.

Metrics. For evaluation, previous efforts (Freitag et al., 2022; Hendy et al., 2023) suggest that utilizing neural network-based metrics, which have demonstrated a high correlation with human evaluation and are resilient to domain shift. Following their recommendations, we employ COMET-22 (Rei et al., 2022a), a reference-based metric that combines direct assessments (DA), sentence-level scores, and word-level tags from Multidimensional Quality Metrics (MQM) error annotations. Additionally, we also evaluate and report the translation quality with BLEU score by tokenized case sensitive SacreBLEU[9].

Implementation Details. In our experiments, we utilized the mt0-large (Muennighoff et al., 2023) model as the backbone for both the fusion model and the student model. For the fusion model training, we use the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 5e-5. We use 4 NVIDIA Tesla V100 GPU cards for the model training, with a batch size of 8 per GPU. We search the value of the margin β in the Eq.10 within the range [1, 10], and the value of 5 is determined based on the model performance on the validation set. We conduct experiments on the validation set using the same approach to search the values of λ, τ, and K in the top-k. In the end, we select $\lambda = 0.6$, $\tau = 2$, and $K = 8$. During the inference of the fusion model, we decode with beam search and set beam size to 4. For all the baseline models, we apply the AdamW optimizer with a learning rate of 6e-5. The batch size is set to 16 per GPU. For all models, we use an early stopping scheduler when there is no improvement on the validation set.

4.2. Main Results

The experiment results are shown in Table 3. For all methods, we adopt a one-to-many approach, training a model in the English->xx directions and a model in the xx->English direction, respectively. Fusion-SL represents using the Fusion model to merge candidate translations and obtain training data directly for training the student model without adding soft labels. We have the following observations: (1) Our TAeKD method achieved the best results in all six language directions on the WMT22 and FLORES-101 test sets, with a 2-point improvement in BLEU score compared to the previous Quality-Aware Sequence-level KD methods. In addition, COMET-22 places more emphasis on evaluating the semantic alignment and fluency of translation results compared to BLEU. In all tests, the student model trained with TAeKD had the highest COMET-22 score, indicating that the translations outputted by the student model trained with TAeKD are superior to other methods in terms of both word-level alignment and semantic-level alignment. (2) Although the Fusion-SL method adopts a similar training strategy to the LaBSE-QA and COMET-QA methods, the experimental comparisons have shown that the Fusion-SL method outperforms these two methods in terms of performance. This highlights the effectiveness of using synthetic training data from fusion models to train the model, and indicates that fine-grained fusion results in higher-quality training data than sentence-level fusion. (3) Compared to Fusion-SL, the TAeKD method that utilizes the soft labels shows a significant improvement, indicating that the soft labels offer a more informative and robust learning signal for the student model, improving its performance and enabling effective knowledge transfer from the teacher model.

4.3. Analysis

In this section, we conduct thorough analyses on the proposed TAeKD method for knowledge distillation from closed-source MNMT Systems.

Ablation Study on Fusion Model. To examine the effectiveness of the proposed fusion model in integrating multi-system translation results, we

[8] https://unbabel.com/research/comet

[9] https://huggingface.co/spaces/evaluate-metric/sacrebleu

Method	WMT22		FLORES-101		WMT22		FLORES-101	
	BLEU	COMET-22	BLEU	COMET-22	BLEU	COMET-22	BLEU	COMET-22
	DE->EN				EN->DE			
ChatGPT-SL	25.48	78.68	33.24	84.25	20.56	79.97	28.19	80.18
MS-SL	24.95	78.40	32.62	83.92	21.30	79.78	28.21	79.86
LaBSE-QA	25.64	78.67	33.44	84.26	21.33	79.98	28.29	80.26
Comet-QA	25.56	78.63	33.21	84.13	21.03	79.82	28.49	80.13
Fusion-SL	25.82	79.09	33.82	84.44	21.70	80.75	29.89	81.06
TAeKD	**26.73**	**79.78**	**35.31**	**85.17**	**22.41**	**81.24**	**31.68**	**82.12**
	CS->EN				EN->CS			
ChatGPT-SL	30.79	77.48	28.52	82.42	18.69	79.21	21.90	81.79
MS-SL	31.05	77.16	28.72	82.25	19.39	79.44	22.47	71.28
LaBSE-QA	31.09	77.36	28.55	82.30	19.31	79.76	22.64	81.68
Comet-QA	31.14	77.37	28.78	82.44	19.57	79.82	22.88	81.93
Fusion-SL	31.96	77.88	29.52	82.75	20.31	80.44	23.44	82.92
TAeKD	**33.68**	**78.86**	**30.62**	**83.65**	**21.66**	**82.00**	**25.19**	**84.73**
	RU->EN				EN->RU			
ChatGPT-SL	29.63	77.86	23.78	80.76	21.54	80.49	18.88	78.57
MS-SL	29.65	77.70	24.84	80.82	22.44	80.30	19.17	78.68
LaBSE-QA	30.50	77.93	24.44	80.86	22.58	80.75	19.58	78.79
Comet-QA	30.62	78.14	24.98	80.99	22.51	80.60	19.34	79.17
Fusion-SL	30.91	78.32	25.17	81.17	22.78	81.19	19.70	79.31
TAeKD	**32.35**	**78.72**	**26.14**	**81.55**	**23.22**	**81.35**	**20.44**	**79.96**

Table 3: The BLEU and COMET-22 scores for English<->German, Russian, and Czech languages are assessed on the WMT22 and FLORES-101 test sets.

Method	en->de	en->cs	en->ru
ChatGPT	53.67	41.78	41.40
MS-Translator	51.30	42.18	40.98
LaBSE	55.86	43.98	43.27
Fusion	**57.76**	**46.49**	**47.37**
w/o QAL	54.98	45.60	45.76

Table 4: Ablation study on the proposed fusion model. The results are based on FuseData test set (w/o indicates without).

λ	0.1	0.2	0.3	0.4
BLEU	26.12	26.34	26.50	26.65
λ	0.5	0.6	0.7	0.8
BLEU	26.70	26.73	26.69	26.65

Table 5: Results for the English->German language direction on the WMT22 test set as λ changes.

conduct ablation experiments on the test set of the FuseData. ChatGPT and MS-Translator represent the direct measurement of the BLEU score of these two systems on the test set. LaBSE represents the BLEU score measured by selecting higher-quality translations from these two systems using the Labse encoding model. Fusion refers to the BLEU score of the synthesized translation, which is obtained through the fine-grained integration of candidate translations using our one-to-many fusion model. w/o QAL represents the result of synthetic translations output by the fusion model trained without a quality assessment learning mechanism. The main results are presented in Table 4. It can be seen that the fusion method outperforms all baselines on most of the datasets. Furthermore, we find an obvious performance drop after removing the quality assessment learning (QAL) mechanism. This indicates that the QAL mechanism enables the model to learn the ability to evaluate the quality of candidate translations, allowing the model to more accurately select segments of higher-quality candidate translations.

Impact of Increasing Training Data for Student Model. We conduct experiments to compare the performance trends of student models trained using different training methods as the training data size increases. As shown in Figure 3, the x-axis represents the varying training data sizes used to train the student model. The ordinate axis represents the average BLEU scores of the student models in the en->de, en->ru, and en-cs directions on the FLORES-101 test set and the WMT22 test set, respectively. As the training data increases, the rate of performance improvement for student models trained using the TAeKD method remains almost unchanged, while the rate of performance improvement for the Comet-QA method gradually slows down. This indicates that compared to other methods, the TAeKD method shows a larger improvement as the data volume increases, highlighting the effectiveness of the proposed TAeKD.

Impact of Top-K and Temperature. In our ex-

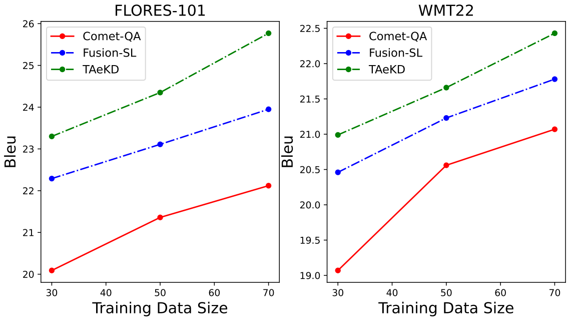


Figure 3: Compare the performance variations of student models trained using different methods as the training data increases. The unit of the horizontal axis is ten thousand.

periments, the student model just matches the top-K output distributions of the teacher model, instead of the full distribution in order to reduce the memory cost. In addition, as shown in Eq.11, we input the top-K output distributions into the softmax function for normalization and use the temperature coefficient τ to adjust the smoothness of the soft labels output by the softmax function. In this section, we conduct experiments on the WMT22 test set in the German->English language direction, with varying values of K (from 2 to 16) and τ (from 1 to 3), to understand their impact on distillation. From the results shown in Figure 4, we see that increasing K from 1 to 8 will improve the BLEU score, while bigger K will bring no gains, even with the full distribution. We conjecture that there may be some noise present in the distribution of lower scores output by the teacher model, which could impact the training of the student model. We also observe that as τ increases, the performance initially improves and reaches its peak, and then it starts to decline.

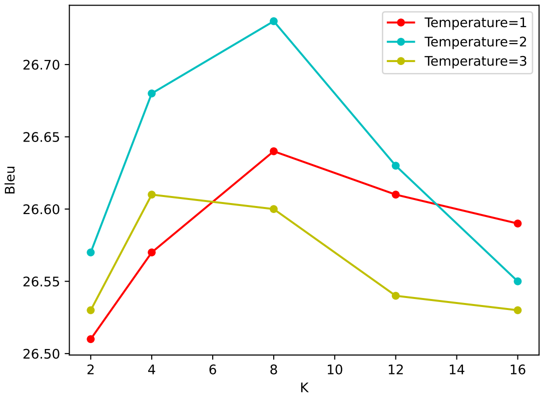


Figure 4: Results for the English->German language direction on the WMT22 test set as K and τ changes.

Impact of Distillation Coefficient λ. For the training objective of TAeKD, we introduce the distillation coefficient λ in Eq.13 to balance the sequence-level KD loss and the word-level KD loss. To analyze the impact of distillation coefficient λ, we show the results of different values of λ on the WMT22 test set in Table 5. We observe that as λ increases, when λ is small, the student does not perform well due to the lack of response-based knowledge of the teacher, and when λ is around 0.6, the student performs best.

5. Related Work

5.1. Multilingual Neural Machine Translation

Multilingual Neural Machine Translation (MNMT) has shown significant promise in developing efficient machine translation systems for numerous languages and enhancing the translation quality for low-resource languages (Johnson et al., 2017; Ha et al., 2016; Dabre et al., 2021). Recent research (Arivazhagan et al., 2019) on factors influencing the performance of MNMT models indicates that the quality and quantity of training data for multilingual models are the primary factors affecting their performance. Moreover, Dabre et al. (2021) demonstrates that MNMT systems tend to generalize better due to their exposure to diverse languages, which leads to improved translation quality compared to bilingual NMT systems. The current most advanced MNMT systems, such as Microsoft Translate, Google Translate (Johnson et al., 2017), and ChatGPT (OpenAI, 2022), have been trained by their respective creators using large-scale, high-quality training data. The performance of these models even surpasses that of previous state-of-the-art bilingual models (Hendy et al., 2023) and exhibits diverse strengths and weaknesses. However, for commercial reasons, these models are typically closed-source, comprising both the model parameters and the training data. To solve this problem, this work aims to study how to efficiently transfer knowledge from these closed-source MNMT systems to a smaller and computationally efficient model.

5.2. Ensemble Learning

Ensemble Learning (Polikar, 2012; Garmash and Monz, 2016; Dong et al., 2020) aims to combine the abilities of different models to compensate for the biases and errors of a single model, thereby achieving better performance. There are plenty of typical ensemble algorithms, such as Adaboost (Freund and Schapire, 1997), Bagging (Breiman, 1996), Stacking (Wolpert, 1992), etc. Weighted average (Wang et al., 2020; Singh and Jaggi, 2020) is an efficient way to fuse several neural networks into a single network (Matena and Raffel, 2022).

Due to the reliance on accessing the weights of each model, these methods cannot be employed to merge the model without exposing its parameters. In contrast, post-hoc ensemble methods (Zhang et al., 2020; Sui et al., 2021) average the output value after inference, thus eliminating the need to access the model parameters. These methods are typically employed in classification tasks, where the final result is determined by the principle of minority submission to the majority based on the classification outcome of each model. Nevertheless, it is difficult to apply these methods to multilingual machine translation tasks, as each sequence involves predicting several tokens. Consequently, we propose a fusion model for post-hoc ensemble in MNMT tasks.

5.3. Knowledge Distillation

Vanilla knowledge distillation (Hinton et al., 2015; Gou et al., 2021) encourages the student model to match the one-hot ground truth and the soft labels provided by the teacher model. Kim and Rush (2016) first refers to vanilla knowledge distillation as word-level KD and further proposes sequence-level KD to address the situation where obtaining soft labels from the teacher model is not possible. To improve the quality of training data, multiple teacher systems can be utilized to translate the same monolingual source data (team et al., 2022). The training data is then selected with the highest sentence similarity score, calculated by a multilingual sentence embedding model (Feng et al., 2022) that uses the source data as a reference. However, the aforementioned sentence-level filtering method lacks the ability to combine the advantages of different candidate translations and cannot provide soft labels. To address these issues, we propose the TAeKD, which uses a fusion model to provide finely integrated translated data and soft labels.

6. Discussion

In this section, we will discuss the significance and ethical implications of our research on knowledge distillation from closed-source MNMT systems.

As mentioned in Section 1, our research has two primary objectives. Firstly, it is to leverage the complementarity of different MNMT systems to generate high-quality training data. Secondly, we aim to compensate for the loss in distillation performance caused by the inability of the teacher model to provide logit scores. Therefore, our proposed method can be applied to address not only the distillation problem of closed-source systems but also other works related to these two issues, that is significance for the MNMT research. This also benefits the development of other research areas within the field of MNMT.

In this line of research, the utilization of results generated by multiple closed-source systems for training students may raise potential ethical concerns, particularly when owners of closed-source systems prohibit the use of results for commercial purposes to enhance other models. Hence, it needs to be emphasized that our method is solely intended for scientific research purposes and cannot be applied for commercial use without the permission from the owners of the closed-source systems. In the future, we hope that our research can promote the study of distilling multiple large models into smaller models, and encourage researchers to develop more high-performance, smaller-sized multilingual translation models.

7. Conclusion

In this paper, we propose TAeKD, a novel teacher assistant enhanced knowledge distillation method for augmenting the capacity of knowledge transfer from closed-source MNMT models. TAeKD adopts a fusion model that can finely fuse candidate translations and provide soft targets to enhance the effectiveness of knowledge distillation. Moreover, a quality assessment learning mechanism is proposed by distinguishing high-quality segments from multiple candidates, thereby enhancing the generalization of the fusion model. To facilitate research on knowledge transfer from MNMT models, we also introduce the FuseData, a benchmark consisting of a blend of translations from multiple closed-source systems. Extensive experiments including ablation studies are carried out to show the effectiveness of TAeKD and its components.

8. Acknowledgements

This work is supported by the National Key Research and Development Program of China (Grant No. 2021ZD0112905), the Major Key Project of PCL (Grant No. PCL2023A09), the National Natural Science Foundation of China (Grant No. 62206140, 62076231), and the China Postdoctoral Science Foundation (Grant No. 2022M711726).

9. Bibliographical References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, MiaXu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *Cornell University - arXiv,Cornell University - arXiv*.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24:123–140.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2021. A survey of multilingual neural machine translation. *ACM Computing Surveys*, page 1–38.

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

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14:241–258.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

MitchellA. Gordon and Kevin Duh. 2019. Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation. *Cornell University - arXiv,Cornell University - arXiv*.

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

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

NunoM. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and Martins. 2023. Hallucinations in large multilingual translation models.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *Cornell University - arXiv,Cornell University - arXiv*.

John Hancock. 2004. *Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient)*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yoon Kim and AlexanderM. Rush. 2016. Sequence-level knowledge distillation. *Cornell University - arXiv,Cornell University - arXiv*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics,Transactions of the Association for Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5191–5198. AAAI Press.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Robi Polikar. 2012. Ensemble learning. *Ensemble machine learning: Methods and applications*, pages 1–34.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Catarina Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. *CoRR*, abs/2010.15535.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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, Berlin, Germany. Association for Computational Linguistics.

Sidak Pal Singh and Martin Jaggi. 2020. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055.

Yi Sui, Ga Wu, and Scott Sanner. 2021. Representer point selection via local jacobian expansion for post-hoc classifier explanation of deep neural networks and ensemble models. *Advances in neural information processing systems*, 34:23347–23358.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *International Conference on Learning Representations,International Conference on Learning Representations*.

Nllb team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. 2020. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*.

Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil Martin Robertson. 2019. Ranked list loss for deep metric learning. *CoRR*, abs/1903.03238.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, pages 11117–11128. PMLR.

Huaiyu Zhu. 1997. On information and sufficiency. *Research Papers in Economics,Research Papers in Economics*.