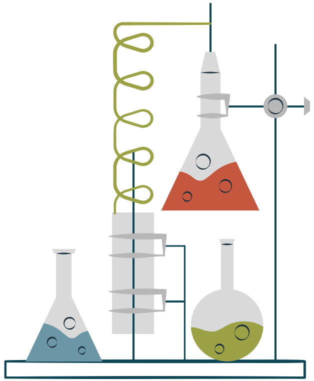
Knowledge Distillation for Machine Translation - KD4MT -



Ona de Gibert
Joseph Attieh

MT Marathon 2024
05.09.2024

GREENNLP

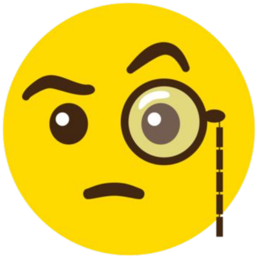
HELSINKI NLP

HPLT

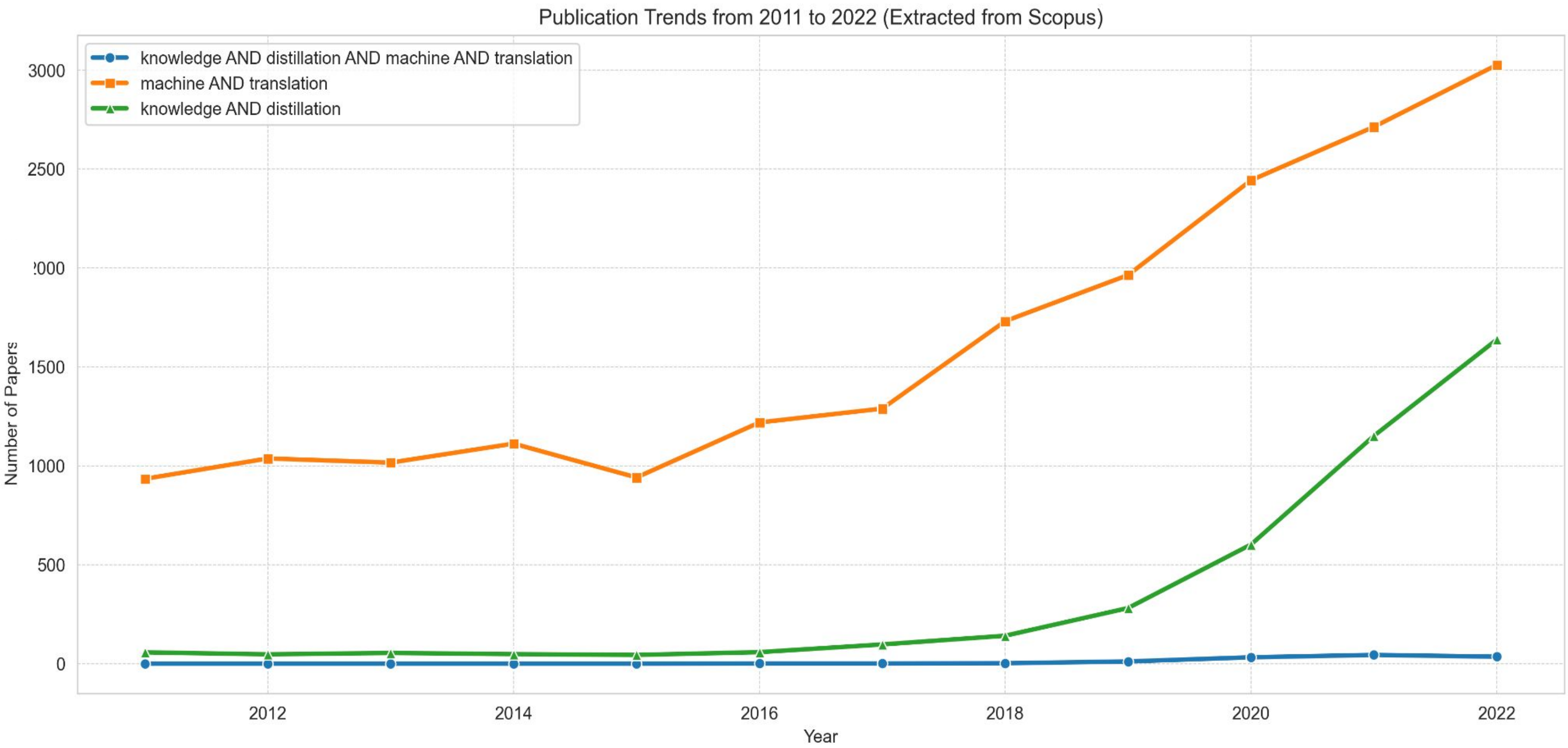
What This Presentation Is About and Is *Not* About

- **Goal**: Provide an overview of the key knowledge distillation methods for Machine Translation
- **What this is not:** Exhaustive

  It's impossible to cover all related papers in one presentation
- **What we cover:**
  Knowledge distillation <u>explicitly</u> applied on Autoregressive NMT models

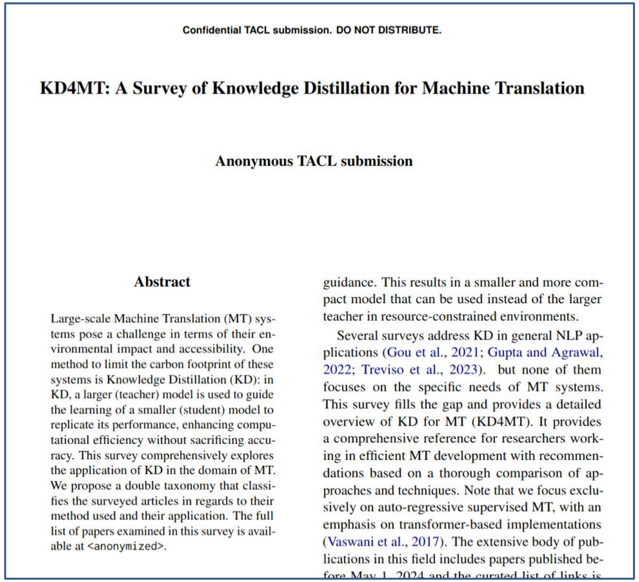


Papers Trend: NMT↗, KD↗, KD for NMT🥴



Theoretical Framework: Survey

- We have conducted a survey of Knowledge Distillation for Machine Translation (KD4MT)



Confidential TACL submission. DO NOT DISTRIBUTE.

KD4MT: A Survey of Knowledge Distillation for Machine Translation

Anonymous TACL submission

Abstract

Large-scale Machine Translation (MT) systems pose a challenge in terms of their environmental impact and accessibility. One method to limit the carbon footprint of these systems is Knowledge Distillation (KD): in KD, a larger (teacher) model is used to guide the learning of a smaller (student) model to replicate its performance, enhancing computational efficiency without sacrificing accuracy. This survey comprehensively explores the application of KD in the domain of MT. We propose a double taxonomy that classifies the surveyed articles in regards to their method used and their application. The full list of papers examined in this survey is available at <anonymized>.

guidance. This results in a smaller and more compact model that can be used instead of the larger teacher in resource-constrained environments.

Several surveys address KD in general NLP applications (Gou et al., 2021; Gupta and Agrawal, 2022; Treviso et al., 2023). but none of them focuses on the specific needs of MT systems. This survey fills the gap and provides a detailed overview of KD for MT (KD4MT). It provides a comprehensive reference for researchers working in efficient MT development with recommendations based on a thorough comparison of approaches and techniques. Note that we focus exclusively on auto-regressive supervised MT, with an emphasis on transformer-based implementations (Vaswani et al., 2017). The extensive body of publications in this field includes papers published before May 1, 2024 and the curated list of links is

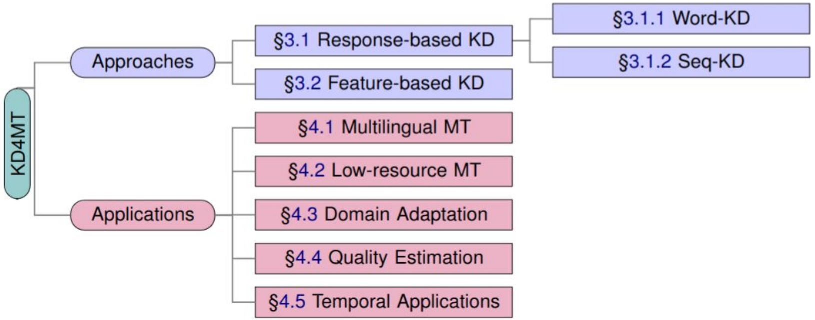


Figure 1: Taxonomy of Knowledge Distillation approaches and applications for Machine Translation.

This talk

I. Introduction
II. Methods
III. Applications
IV. KD4MT @ Helsinki-NLP

Introduction

Introduction

Rapid Advances in NLP and MT: The Trend Toward Larger Models

- Increasing model size:
  - Better translation quality
  - Greater multilingual capabilities
  - Increased robustness
  - **Best results with:** Ensemble Models, Mixture of Experts, or Large Networks
- Example: NLLB can translate across 202 languages

System	Size	X-en COMET	X-en BLEU	X-zh COMET	X-zh BLEU	X-de COMET	X-de BLEU
Encoder-Decoder Models							
M2M-100* (Fan et al., 2021)	418M	68.47	21.19	62.15	10.34	60.19	14.25
M2M-100* (Fan et al., 2021)	1.2B	73.06	26.26	67.91	12.94	67.78	19.33
M2M-100* (Fan et al., 2021)	12B	74.45	28.01	69.27	13.35	70.17	21.31
Lego-MT* (Yuan et al., 2023a)	1.2B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NLLB-200 (Team et al., 2022)	1.3B	<u>84.22</u>	<u>38.60</u>	76.75	15.27	<u>79.50</u>	25.71
MADLAD-400 (Kudugunta et al., 2024)	7B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Aya-101 (Üstün et al., 2024)	13B	80.72	31.92	<u>78.51</u>	<u>22.49</u>	77.37	15.43
LLM Based Decoder-Only Models							
LLaMA2 (Touvron et al., 2023b)	7B	55.46	11.80	43.50	0.55	43.10	3.22
LLaMA2 (Touvron et al., 2023b)	13B	38.25	0.75	37.06	0.22	31.73	0.25
LLaMA3 (AI@Meta, 2024)	8B	67.66	19.81	42.52	1.37	49.42	6.61
LLaMA2-Alpaca (Taori et al., 2023)	7B	65.85	16.44	56.53	4.46	56.76	9.01
LLaMA2-Aplaca (Taori et al., 2023)	13B	68.72	19.69	64.46	8.80	62.86	12.57
LLaMA3-Alpaca (Taori et al., 2023)	8B	77.43	26.55	73.56	13.17	71.59	16.82
PolyLM (Wei et al., 2023)	13B	50.98	7.75	42.60	1.20	43.95	3.69
Yayi2 (Luo et al., 2023)	30B	68.06	19.37	57.81	6.07	53.82	5.62
TowerInstruct (Alves et al., 2024)	7B	65.37	18.87	64.26	10.37	60.73	12.81
Aya-23 (Aryabumi et al., 2024)	8B	67.53	20.57	66.11	11.20	63.09	14.09
Qwen2-Instruct (Bai et al., 2023)	7B	73.25	19.04	72.52	13.52	64.61	11.33
ChineseLLaMA2-Alpaca (Cui et al., 2024)	7B	-	-	55.06	6.15	-	-
LLaMAX2-Alpaca	7B	80.55	30.63	75.52	13.53	74.47	19.26
LLaMAX3-Alpaca	8B	**81.28**	**31.85**	**78.34**	**16.46**	**76.23**	**20.64**

[Lu et. al (July 2024)]

Introduction

- **Example:** NLLB can translate across 202 languages but raises significant concerns
- **Challenges of Large-Scale Models:**
  - **Accessibility Issues:**
    - Limited computational resources to train and run these models
    - Difficulty deploying on edge devices

Constraints on Model Scale Our research is confined to language models of a moderate size, specifically those with $7B$ parameters. This limitation is due to the constraints of our available resources. Consequently, it is crucial to acknowledge that the outcomes of our study might vary if conducted with larger models.

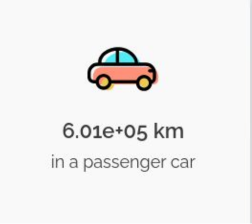
Wu et al. (June 2024)

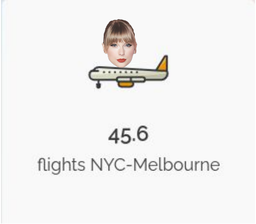
Introduction

- **Example:** NLLB can translate across 202 languages but raises significant concerns
- **Challenges of Large-Scale Models:**
  - **Accessibility Issues:**
    - Limited computational resources to train and run these models
    - Difficulty deploying on edge devices
  - **Environmental Impact:**
    - Higher energy consumption
    - Higher carbon footprint

	Time (h)	Power Consumption (W)	Carbon Emitted (tCO_2eq)
Data Mining	108,366	400	17.55
Backtranslation	18,000	300	2.17
Modeling	196,608	400	31.74
Final Ablations	224,000	400	36.17
Evaluations	51,200	400	8.26
NLLB-200	51,968	400	8.39
Total			104.31

[NLLB Team (July 2022)]



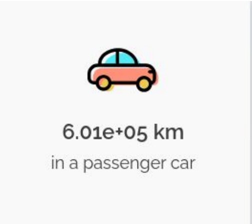


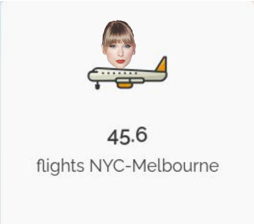
Introduction

- **Example:** NLLB can translate across 202 languages but raises significant concerns
- **Challenges of Large-Scale Models:**
  - **Accessibility Issues:**
    - Limited computational resources to train and run these models
    - Difficulty deploying on edge devices
  - **Environmental Impact:**
    - Higher energy consumption
    - Higher carbon footprint
- How can we reduce the size of models while maintaining their high level of performance?

	Time (h)	Power Consumption (W)	Carbon Emitted (tCO_2eq)
Data Mining	108,366	400	17.55
Backtranslation	18,000	300	2.17
Modeling	196,608	400	31.74
Final Ablations	224,000	400	36.17
Evaluations	51,200	400	8.26
NLLB-200	51,968	400	8.39
Total			104.31

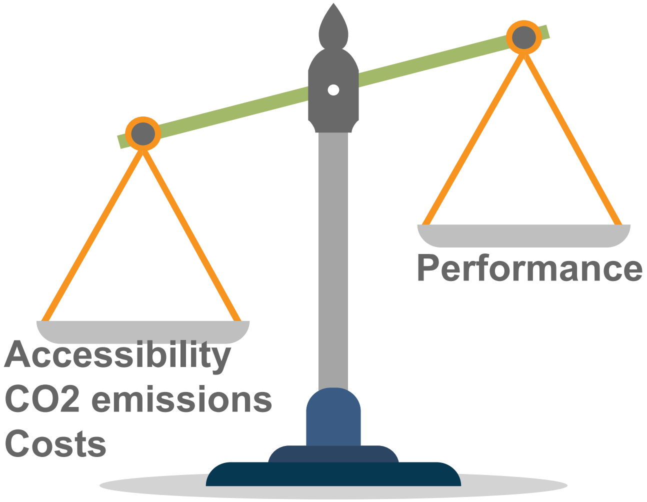
[NLLB Team (July 2022)]





Balancing Model Size and Performance

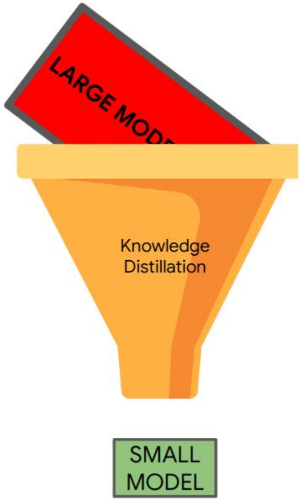
How can we reduce the size of models
<u>without major drop in performance?</u>

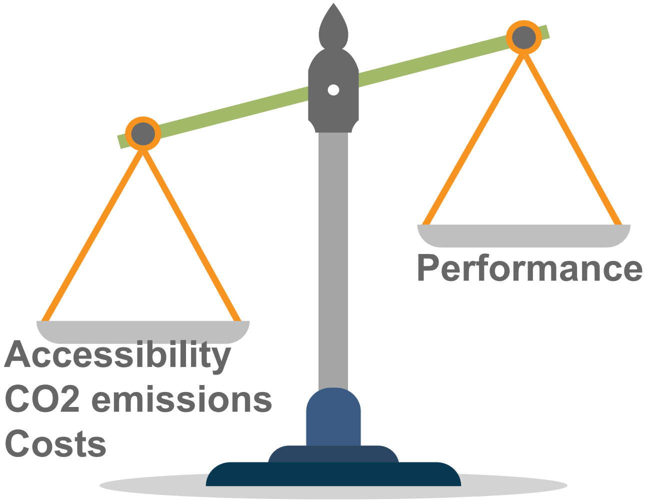


Balancing Model Size and Performance

How can we reduce the size of models
without major drop in performance?

Knowledge Distillation



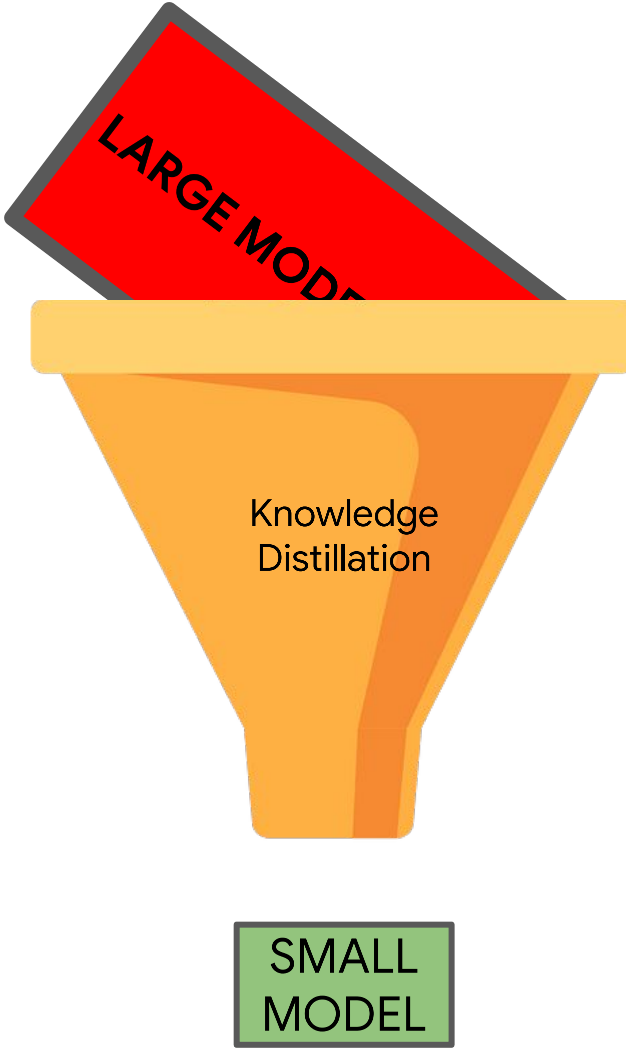


What is Knowledge Distillation?

- Transferring the knowledge from a (set of) **large** model(s) to a **smaller** model w/o significant loss in performance.

- The small model is a **student** that learns from the large **teacher** model by imitating the teacher predictions.

- Advantages of having a student model:
    - reduced computational demands
    - maintaining performance in resource-constrained environments.

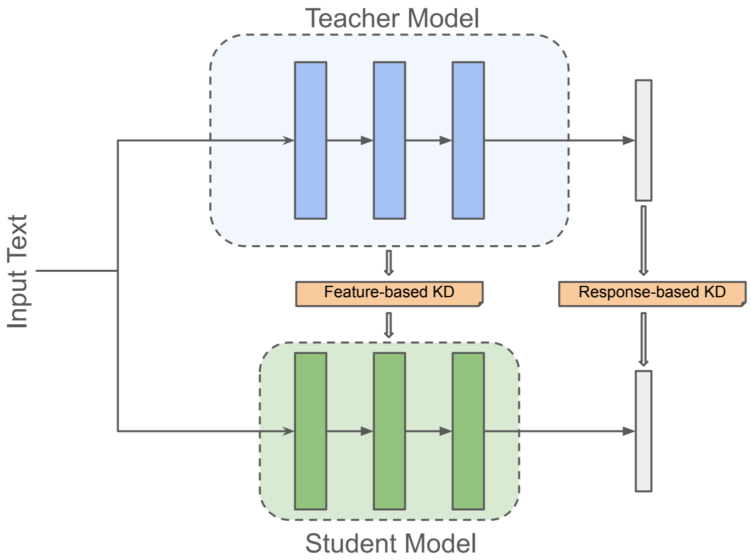


How is KD performed for NMT models?

How is KD performed for NMT models?

Two families of methods

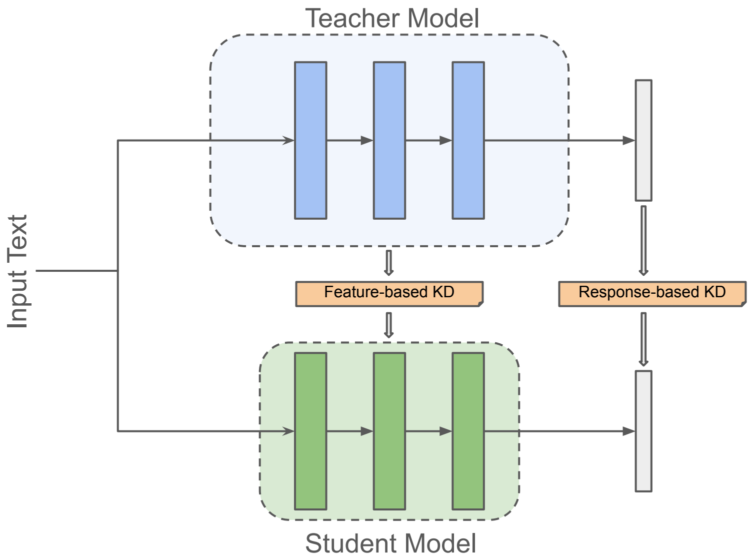
- **Response-Based Methods**
  - Focus on the final predictions of the teacher model
  - <u>Examples</u>: Word-Level KD, Sequence-Level KD
- **Feature-Based Methods**
  - Transfer knowledge from intermediate layers of the teacher model to the student model
  - <u>Examples</u>: Layer-wise supervision, weight distillation



How is KD performed for NMT models?

Two families of methods

- **Response-Based Methods**
    - Focus on the final predictions of the teacher model
    - Examples: Word-Level KD, Sequence-Level KD
- **Feature-Based Methods**
    - Transfer knowledge from intermediate layers of the teacher model to the student model
    - Examples: Layer-wise supervision, weight distillation



Response-based Methods

Word-level KD

Word-level KD

Word-level KD

- **Objective:** The student model is trained to output a similar distribution as the teacher model for every token.

Word-level KD

- **Objective:** The student model is trained to output a similar distribution as the teacher model for every token.
- **Method:** The loss between the student model and the teacher probability distribution is minimized, instead of using the observed data directly.

Auto-regressive Negative Log-Likelihood (NLL) Loss:

$$L_{NLL} = -\sum_{j=1}^{|J|} \sum_{k=1}^{|V|} \mathbb{1}\{t_j = k\} \, log \; p_\theta(t_j = k | s, t_{<j})$$

Word-level KD

- **Objective:** The student model is trained to output a similar distribution as the teacher model for every token.
- **Method:** The loss between the student model and the teacher probability distribution is minimized, instead of using the observed data directly.

Auto-regressive Negative Log-Likelihood (NLL) Loss:

$$L_{NLL} = -\sum_{j=1}^{|J|}\sum_{k=1}^{|V|} \mathbb{1}\{t_j = k\} \, log \; p_\theta(t_j = k|s, t_{<j})$$

Having access to a teacher distribution

$$L_{WORD-KD} = -\sum_{j=1}^{|J|}\sum_{k=1}^{|V|} q(t_j = k \,|s, t_{<j}) \; log \; p_\theta(t_j = k|s, t_{<j})$$

compares the student predicted probability distribution with the teacher's (~data distr)

Word-level KD

- **Objective:** The student model is trained to output a similar distribution as the teacher model for every token.
- **Method:** The loss between the student model and the teacher probability distribution is minimized, instead of using the observed data directly.

  Auto-regressive Negative Log-Likelihood (NLL) Loss:

  $$L_{NLL} = -\sum_{j=1}^{|J|}\sum_{k=1}^{|V|} \mathbb{1}\{t_j = k\} \, log \, p_\theta(t_j = k|s, t_{<j})$$

  Having access to a teacher distribution

  $$L_{WORD-KD} = -\sum_{j=1}^{|J|}\sum_{k=1}^{|V|} q(t_j = k \,|s, t_{<j}) \, log \, p_\theta(t_j = k|s, t_{<j})$$

  compares the student predicted probability distribution with the teacher's (~data distr)

- **Final Loss:** $$\mathcal{L}(\theta;\theta_T) = (1-\alpha)\mathcal{L}_{NLL}(\theta) + \alpha\mathcal{L}_{\text{WORD-KD}}(\theta;\theta_T)$$

Word-level KD

- **Objective:** The student model is trained to output a similar distribution as the teacher model for every token.
- **Method:** The loss between the student model and the teacher probability distribution is minimized, instead of using the observed data directly.

  Auto-regressive Negative Log-Likelihood (NLL) Loss:

  $$L_{NLL} = -\sum_{j=1}^{|J|}\sum_{k=1}^{|V|} \mathbb{1}\{t_j = k\}\, log\; p_\theta(t_j = k | s,\, t_{<j})$$

  Having access to a teacher distribution

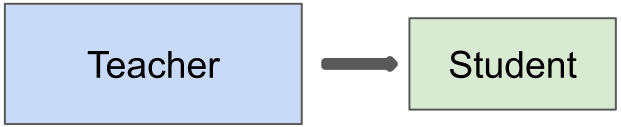
  $$L_{WORD-KD} = -\sum_{j=1}^{|J|}\sum_{k=1}^{|V|} q(t_j = k\; | s,\, t_{<j})\; log\; p_\theta(t_j = k | s,\, t_{<j})$$

  compares the student predicted probability distribution with the teacher's (~data distr)

- **Final Loss:** $$\mathcal{L}(\theta;\theta_T) = (1-\alpha)\mathcal{L}_{NLL}(\theta) + \alpha\mathcal{L}_{\text{WORD-KD}}(\theta;\theta_T)$$

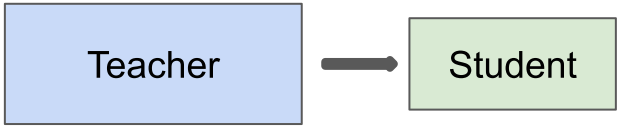
- **Practical Implementation:** At each time step, Word-KD computes the predictions from both the student and the teacher, and then calculates the relevant losses.

Variants of Word-KD



Problem: Word-KD performance result in a performance drop between teacher and student

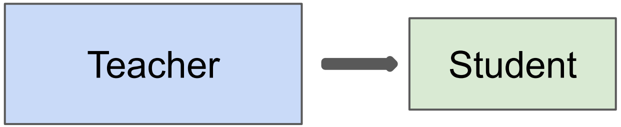
Variants of Word-KD



Problem: Word-KD performance result in a performance drop between teacher and student

Why so?

Variants of Word-KD

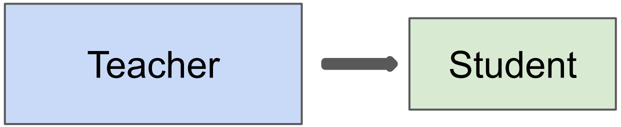


Problem: Word-KD performance result in a performance drop between teacher and student

Why so?

Capacity Gap Problem

Variants of Word-KD



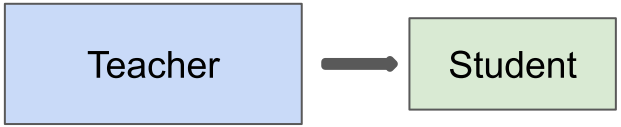
Problem: Word-KD performance result in a performance drop between teacher and student

Why so?

Capacity Gap Problem

- when the size gap between the teacher and student increases, training the student using KD becomes more difficult
- size gap → performance gap

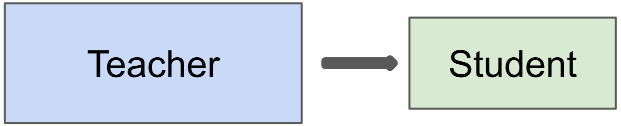
Variants of Word-KD



Problem: Word-KD performance result in a performance drop between teacher and student

Objective: Refine the process of knowledge transfer

Variants of Word-KD

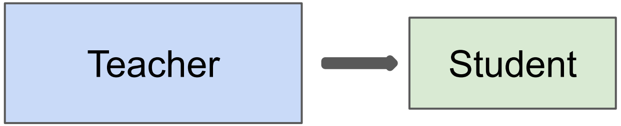


Problem: Word-KD performance result in a performance drop between teacher and student

Objective: Refine the process of knowledge transfer

Key Methods:

Variants of Word-KD



Problem: Word-KD performance result in a performance drop between teacher and student

Objective: Refine the process of knowledge transfer

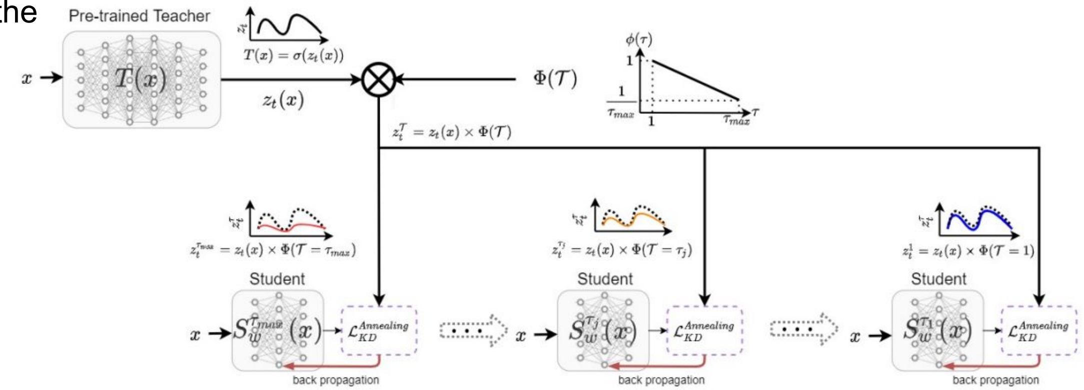
Key Methods:

1. Annealing Distillation (Jafari et al., 2021)

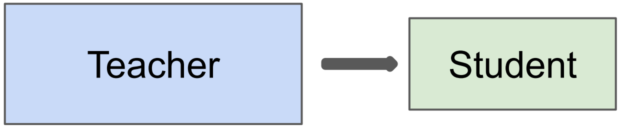
- Incrementally introduce soft targets from the teacher to the student at varying temperatures using MSE loss
- Smooths the knowledge transfer process, bridging the capacity gap

$$\mathcal{L}_{\text{KD}}^{\text{Annealing}}(i) = ||z_s(x) - z_t(x) \times \Phi(\mathcal{T}_i)||_2^2$$

$$\Phi(\mathcal{T}) = 1 - \frac{\mathcal{T} - 1}{\tau_{\max}}, 1 \leq \mathcal{T} \leq \tau_{\max}, \mathcal{T} \in \mathbb{N}$$



Variants of Word-KD



Problem: Word-KD performance result in a performance drop between teacher and student

Objective: Refine the process of knowledge transfer

Key Methods:

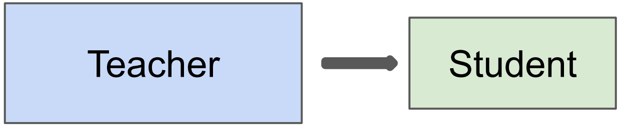
2. Selective Distillation (Wang et al., 2021)

- Distilling knowledge from all samples is not always optimal
- Word CE measures how the student model agrees with the golden label
- Words with large CE are more difficult to learn and get extra supervision signal from teacher (i.e., distillation)

$$\mathcal{L}(\theta;\theta_T) = (1-\alpha)\mathcal{L}_{\mathrm{NLL}}(\theta) + \alpha\mathcal{L}_{\mathrm{KD}}(\theta;\theta_T)$$

$$\mathcal{L}_{kd} = \begin{cases} -\sum_{k=1}^{|V|} q(y_k)\cdot\log p(y_k), & y \in \mathcal{S}_{Hard} \\ 0 & , y \in \mathcal{S}_{Easy} \end{cases}$$

Variants of Word-KD



Problem: Word-KD performance result in a performance drop between teacher and student

Objective: Refine the process of knowledge transfer

Key Methods:

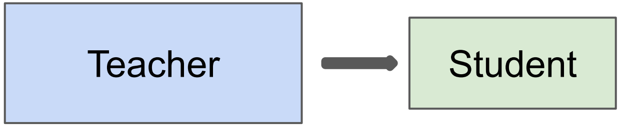
2. Selective Distillation (Wang et al., 2021)

- Distilling knowledge from all samples is not always optimal
- Word CE measures how the student model agrees with the golden label
- Words with large CE are more difficult to learn and get extra supervision signal from teacher (i.e., distillation)

Models	En-De	Δ
Transformer	27.29	ref
Word-KD	28.14	+0.85
Batch-level Selection	28.42*	+1.13
Global-level Selection	**28.57*†**	**+1.28**

Table 2: BLEU scores (%) on WMT'14 English-German (En-De) task. Δ shows the improvement compared to Transformer (Base). '*': significantly ($p < 0.01$) better than Transformer (Base). '†': significantly ($p < 0.05$) better than the Word/Seq-KD models.

Variants of Word-KD



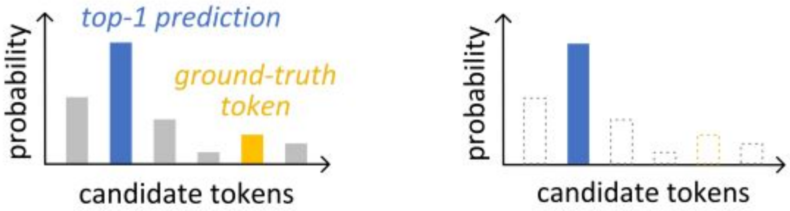
Problem: Word-KD performance result in a performance drop between teacher and student

Objective: Refine the process of knowledge transfer

Key Methods:

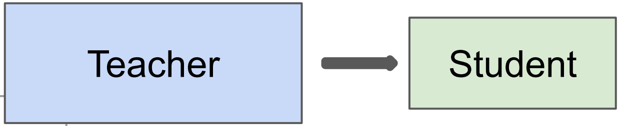
3. Top Information Enhanced-KD (Zhang et al., 2023a)

- The knowledge transferred during KD actually comes from the top-1 predictions of the teacher
- Word-level KD lacks specialized learning of that information
- TIEKD enforces the student model to learn the top-1 information from the teacher by ranking the teacher's top-1 predictions as its own top-1 predictions



(a) vanilla word-level KD (b) *w/o* correlation info

Variants of Word-KD



Problem: Word-KD performance result in a performance drop between teacher and student

Objective: Refine the process of knowledge transfer

Key Methods:

Methods	WMT'14 En-De		WMT'14 En-Fr		WMT'16 En-Ro	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
Student (*Transformer$_{base}$*)	$27.42_{\pm 0.01}$	$48.11_{\pm 1.04}$	$40.97_{\pm 0.14}$	$62.19_{\pm 0.11}$	$33.59_{\pm 0.15}$	$50.96_{\pm 0.43}$
+ Word-KD (Kim and Rush, 2016)	$28.03_{\pm 0.10}$	$51.59_{\pm 0.23}$	$41.10_{\pm 0.11}$	$63.81_{\pm 0.14}$	$33.77_{\pm 0.01}$	$53.15_{\pm 0.26}$
+ Annealing KD (Jafari et al., 2021)	$27.91_{\pm 0.10}$	$51.58_{\pm 0.03}$	$41.20_{\pm 0.13}$	$63.59_{\pm 0.09}$	$33.67_{\pm 0.09}$	$52.22_{\pm 1.02}$
+ Selective-KD (Wang et al., 2021)	$28.24_{\pm 0.21}$	$52.15_{\pm 0.42}$	$41.25_{\pm 0.04}$	$64.24_{\pm 0.01}$	$33.74_{\pm 0.02}$	$53.05_{\pm 0.28}$
+ TIE-KD (ours)	$\mathbf{28.46}^{*}_{\pm 0.01}$	$\mathbf{52.63}^{*}_{\pm 0.09}$	$\mathbf{41.57}^{*}_{\pm 0.08}$	$\mathbf{65.06}^{*}_{\pm 0.44}$	$\mathbf{34.70}^{*}_{\pm 0.07}$	$\mathbf{55.76}^{*}_{\pm 0.21}$
Teacher (*Transformer$_{big}$*)	*28.81*	*53.20*	*42.98*	*69.58*	*34.70*	*57.04*

Table 6: BLEU scores (%) and COMET (Rei et al., 2020) scores (%) on three translation tasks. Results with † are taken from the original papers. Others are our re-implementation results using the released code with the same setting in Sec.5.2 for a fair comparison. We report average results over 3 runs with random initialization. Results with $*$ are statistically (Koehn, 2004) better than the vanilla Word-KD with $p < 0.01$.

Response-based Methods

Sequence-level KD

Sequence-level KD

Sequence-level KD

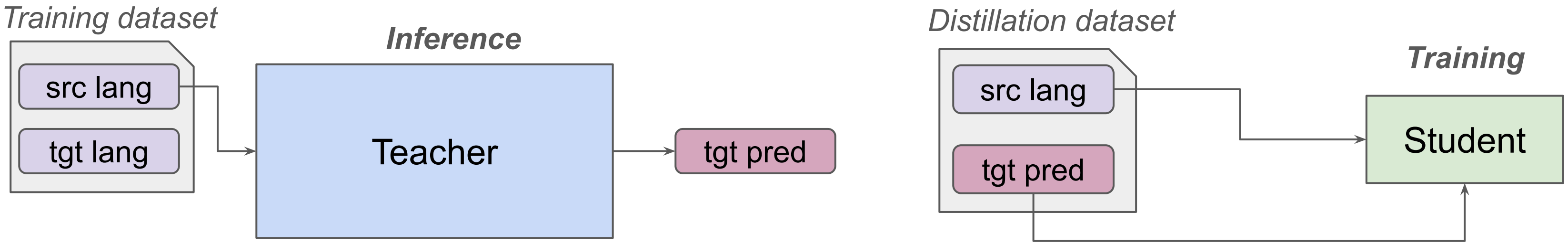
- **Objective:** The student model is trained to mimic the behavior of the teacher model at the sentence level.

Sequence-level KD

- **Objective:** The student model is trained to mimic the behavior of the teacher model at the sentence level.
- **Method:**
  - Instead of minimizing word-level CE, minimize CE between sequence distributions
  - This involves matching the predicted sequence of the student to the teacher sequence.

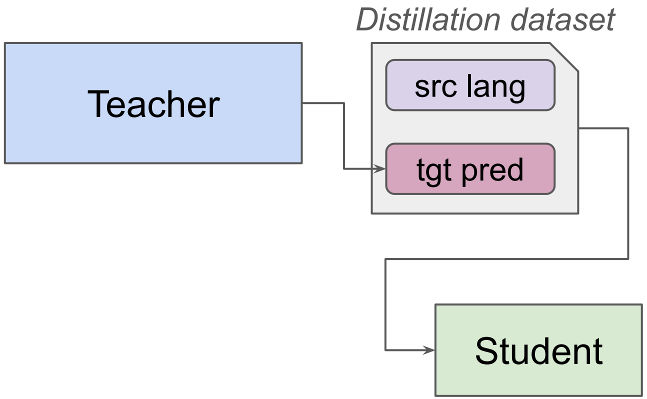
Sequence-level KD

- **Objective:** The student model is trained to mimic the behavior of the teacher model at the sentence level.
- **Method:**
  - Instead of minimizing word-level CE, minimize CE between sequence distributions
  - This involves matching the predicted sequence of the student to the teacher sequence.
- **Practical Implementation:** Seq-KD reduces to a two-step procedure



Variants of Seq-KD

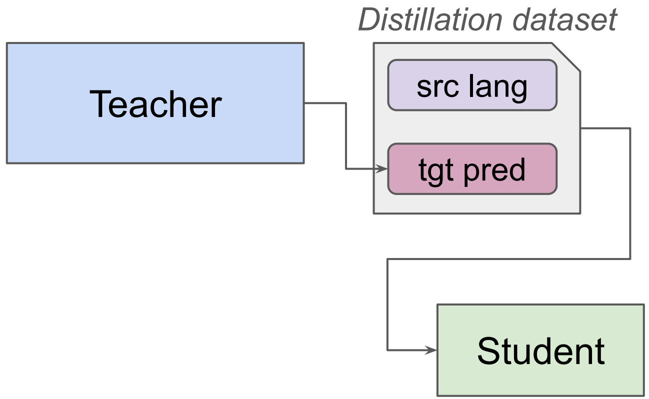
Problem: Can we optimize the construction of the Distillation Set?



Variants of Seq-KD

Problem: Can we optimize the construction of the Distillation Set?

Objective: Enhance the quality of the distillation set by selecting or modifying the data used for training the student model.

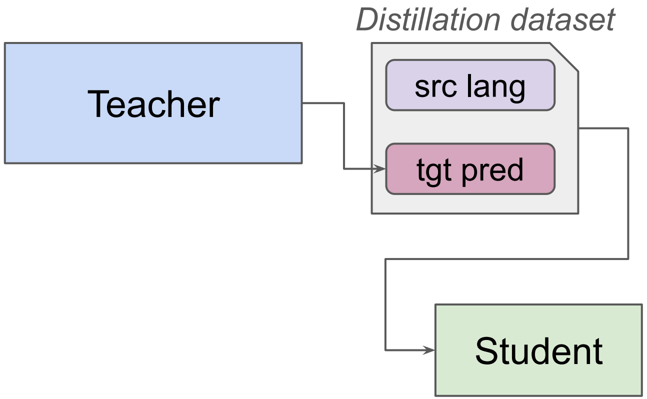


Variants of Seq-KD

Problem: Can we optimize the construction of the Distillation Set?

Objective: Enhance the quality of the distillation set by selecting or modifying the data used for training the student model.

Key Methods:



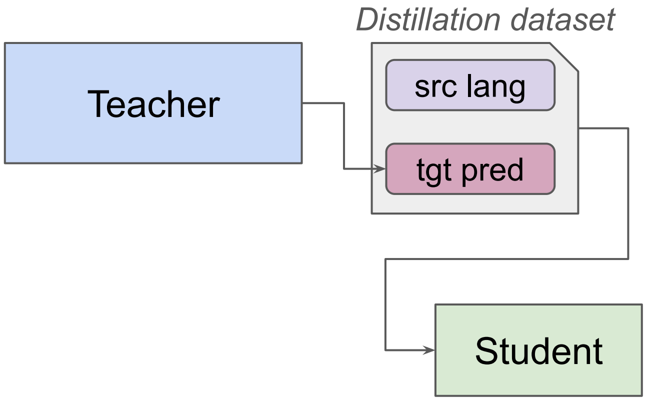
Variants of Seq-KD

Problem: Can we optimize the construction of the Distillation Set?

Objective: Enhance the quality of the distillation set by selecting or modifying the data used for training the student model.

Key Methods:

- **Sequence-Level Interpolation (Kim and Rush, 2016):**
  - Uses beam search to generate multiple candidate translations.
  - Selects the best candidate based on similarity to the training target sequence using sentence-level BLEU.



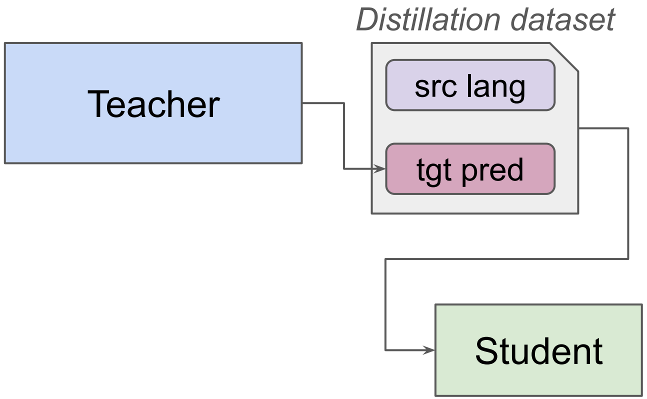
Variants of Seq-KD

Problem: Can we optimize the construction of the Distillation Set?

Objective: Enhance the quality of the distillation set by selecting or modifying the data used for training the student model.

Key Methods:

- **Sequence-Level Interpolation (Kim and Rush, 2016):**
  - Uses beam search to generate multiple candidate translations.
  - Selects the best candidate based on similarity to the training target sequence using sentence-level BLEU.
- **Noise Filtering and Replacement (Zhang et al., 2018):**
  - Filters and replaces noisy translations in the distillation set.
  - Noisy translations are considered as the ones that are not similar to their source sentences, detected using (Pham et al., 2018)



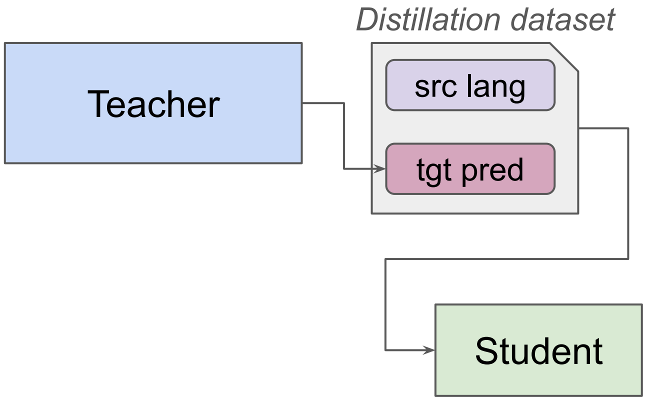
Variants of Seq-KD

Problem: Can we optimize the construction of the Distillation Set?

Objective: Enhance the quality of the distillation set by selecting or modifying the data used for training the student model.

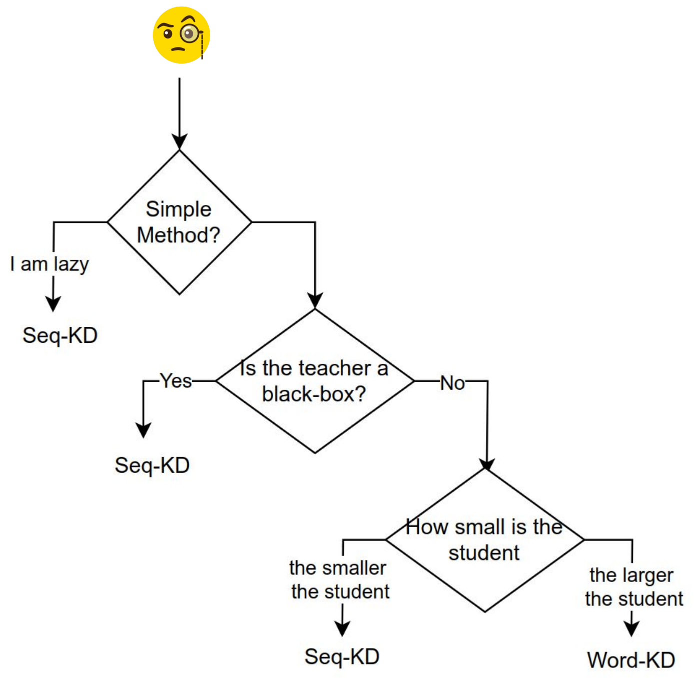
Key Methods:

- **Sequence-Level Interpolation (Kim and Rush, 2016):**
  - Uses beam search to generate multiple candidate translations.
  - Selects the best candidate based on similarity to the training target sequence using sentence-level BLEU.
- **Noise Filtering and Replacement (Zhang et al., 2018):**
  - Filters and replaces noisy translations in the distillation set.
  - Noisy translations are considered as the ones that are not similar to their source sentences, detected using (Pham et al., 2018)
- **MT-PATCHER (Li et al., 2024):**
  - Utilizes LLMs to identify student errors and design corrective training samples.

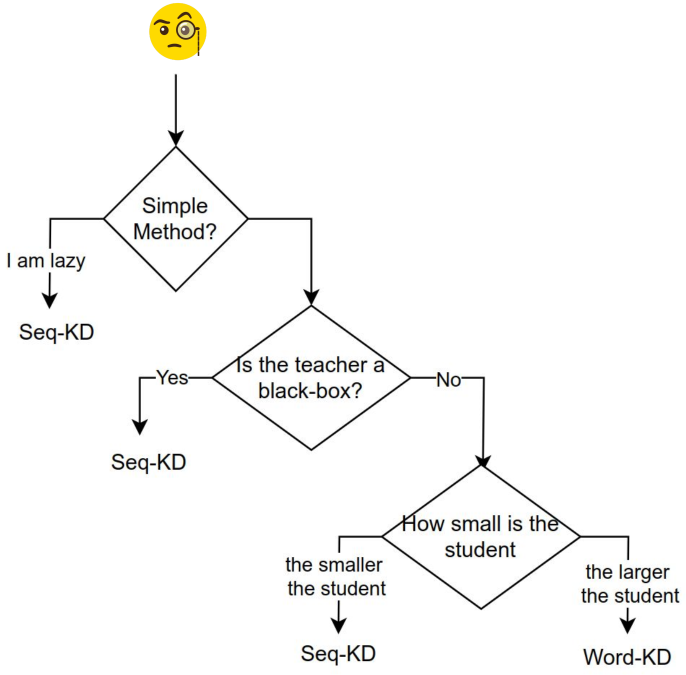


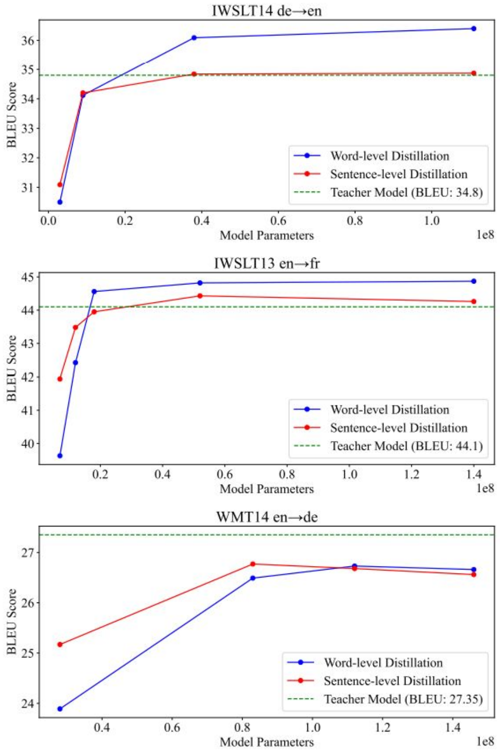
How to choose the appropriate method for distillation?

How to choose the appropriate method for distillation?



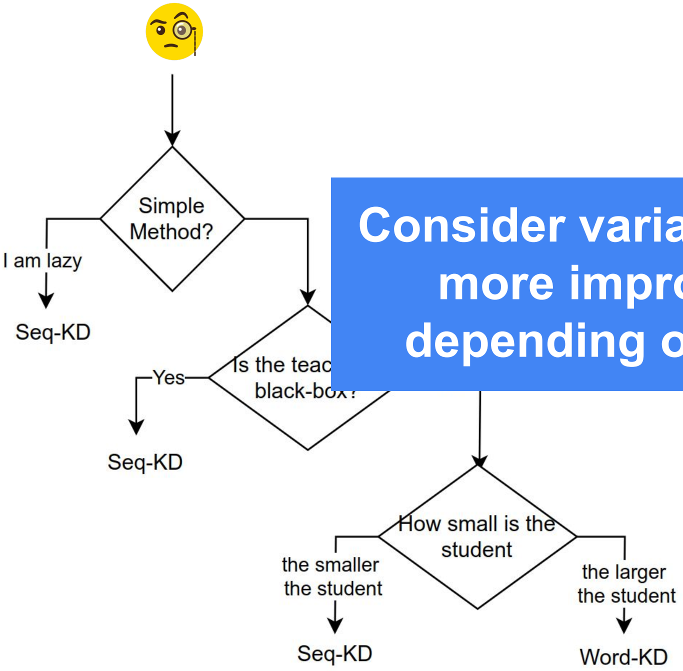
How to choose the appropriate method for distillation?



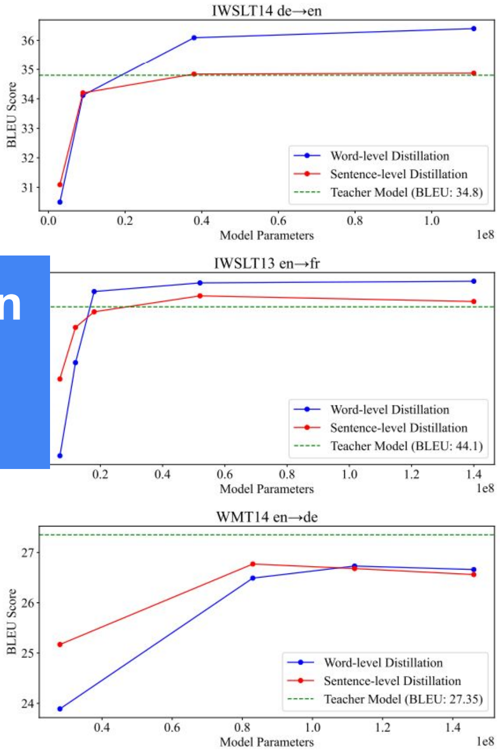


(Wei et al., 2023)

How to choose the appropriate method for distillation?



Consider variants for even more improvement, depending on the data



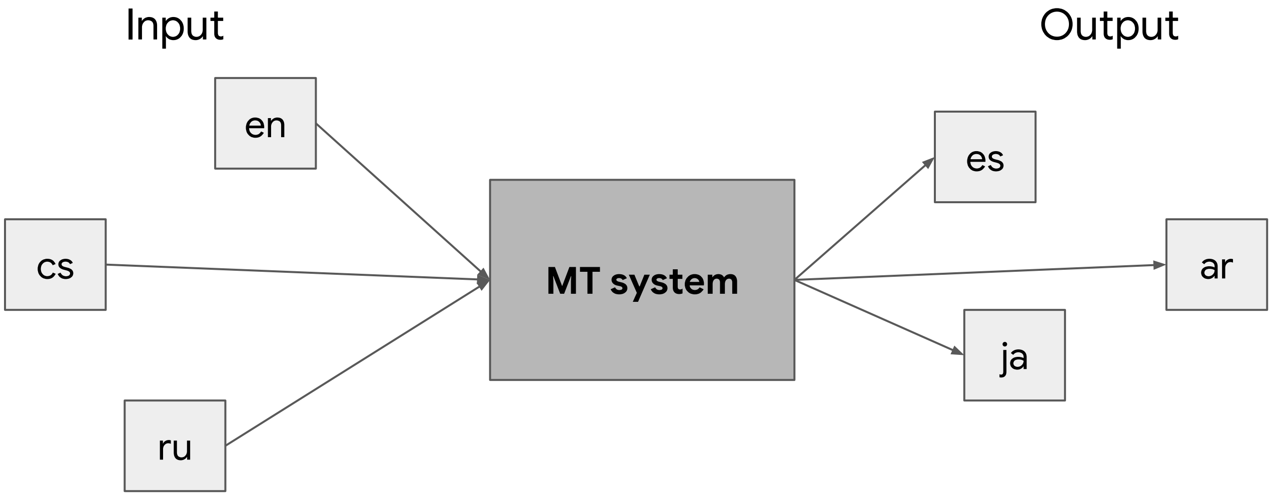
Applications

Applications

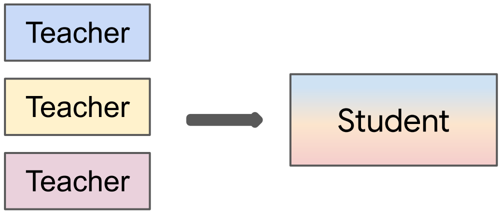
Multilingual MT
Massively Multilingual MT
Low-resource MT

Multilingual MT

What a single MT model to translate from or into multiple languages (Dabre et al. 2020).



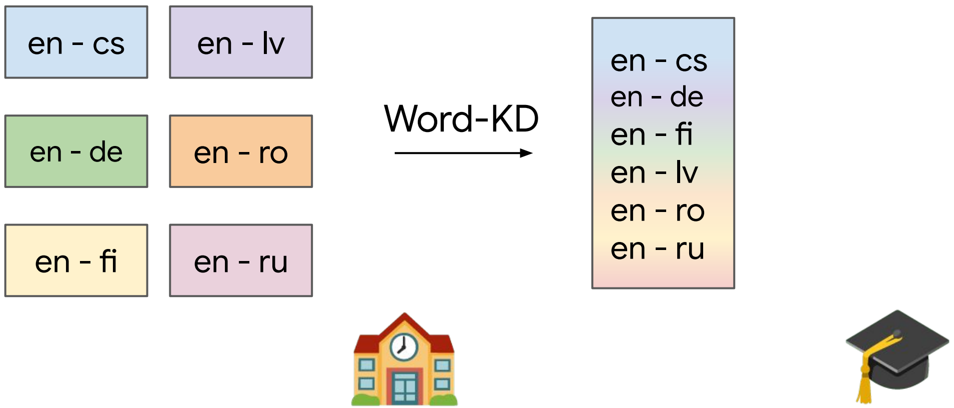
Multilingual MT



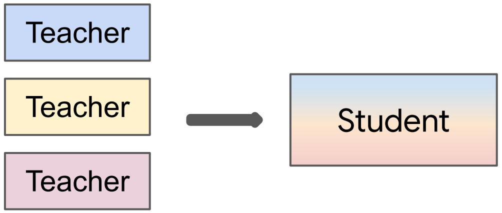
Key Studies

[1] Tan et al. (2019)

- **Selective KD**: distill only when teacher surpasses student



Multilingual MT

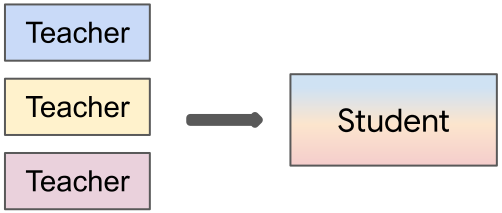


Key Studies

[1] Tan et al. (2019)

- **Selective KD**: distill only when teacher surpasses student

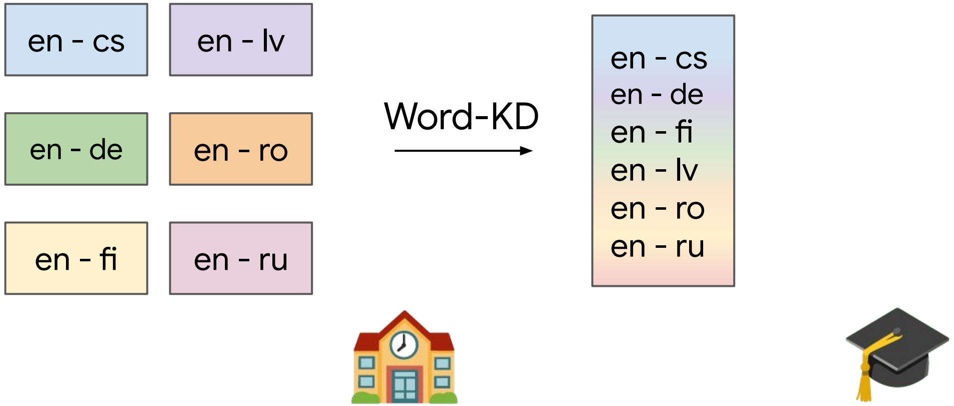
Multilingual MT



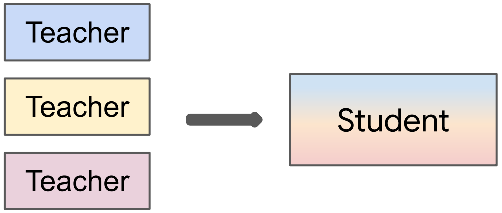
Key Studies

[1] Tan et al. (2019)

- **Selective KD**: distill only when teacher surpasses student
- **Top-k KD**: load the top-K probabilities of the distribution into memory → Top-8
- **Back-distillation**: use the distilled model as a teacher

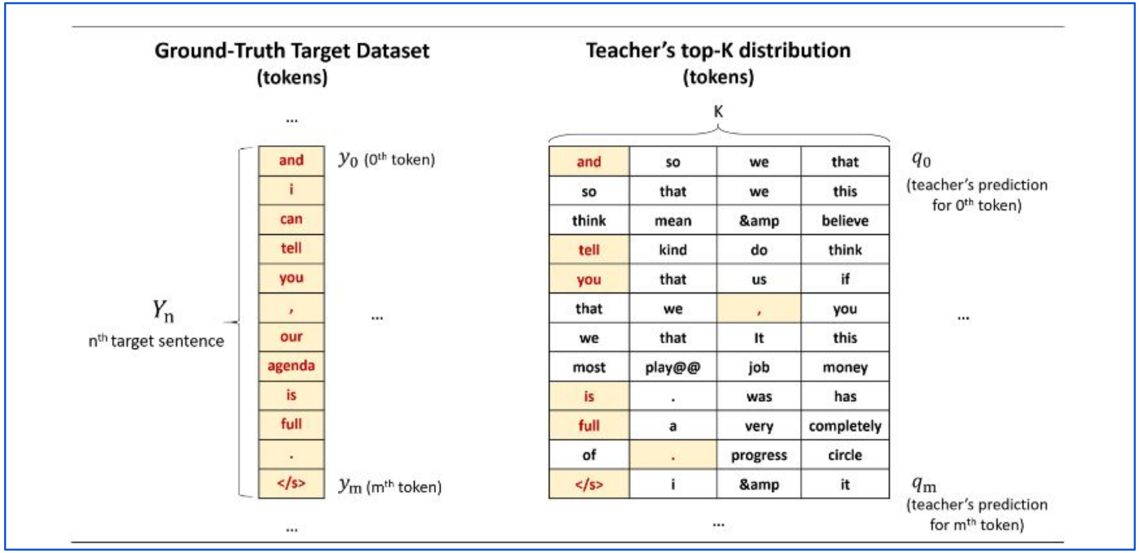


Multilingual MT

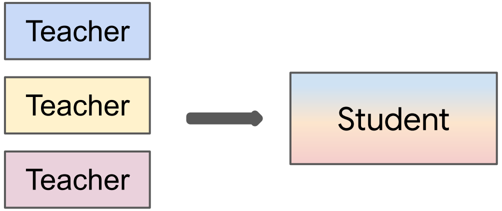


Key Studies

Problem: Top-K KD: the distributions do not always include the ground truth.



Multilingual MT

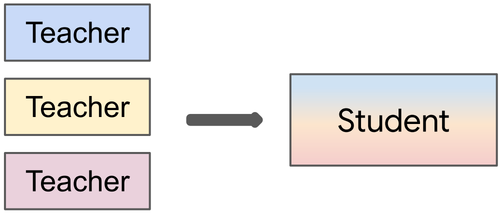


Key Studies

[2] Do and Lee (2023)

- **Target-oriented KD**: penalty for samples that lack the ground truth in their top-K.

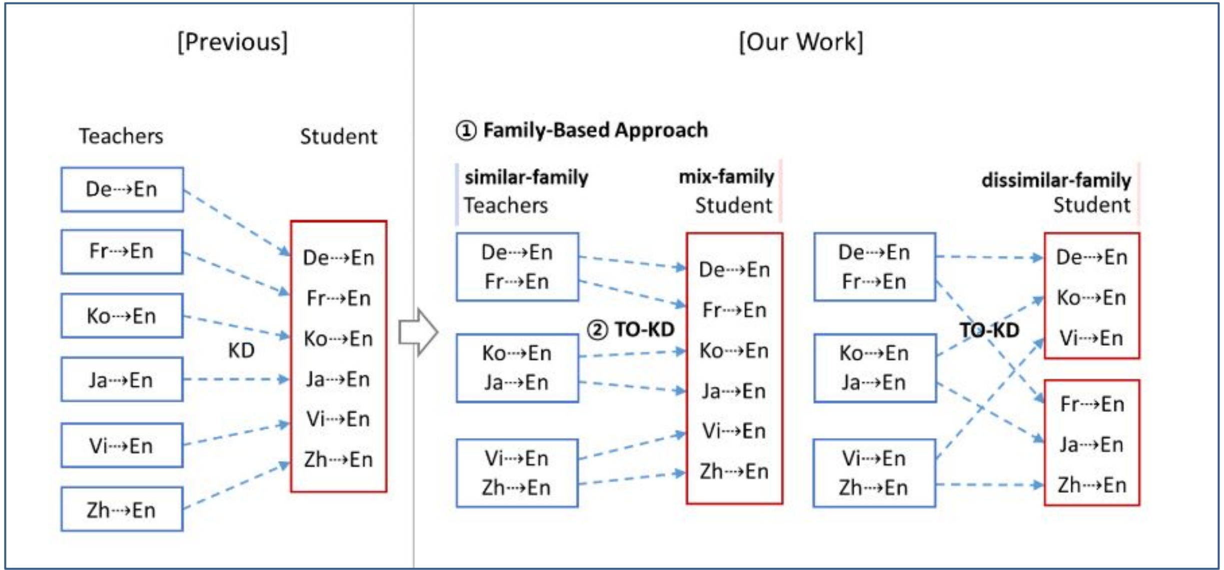
Multilingual MT



Key Studies

[2] Do and Lee (2023)

- **Target-oriented KD**: penalty for samples that lack the ground truth in their top-K.
- Family-based KD (Sun et al., 2020)



Multilingual MT

Takeaways

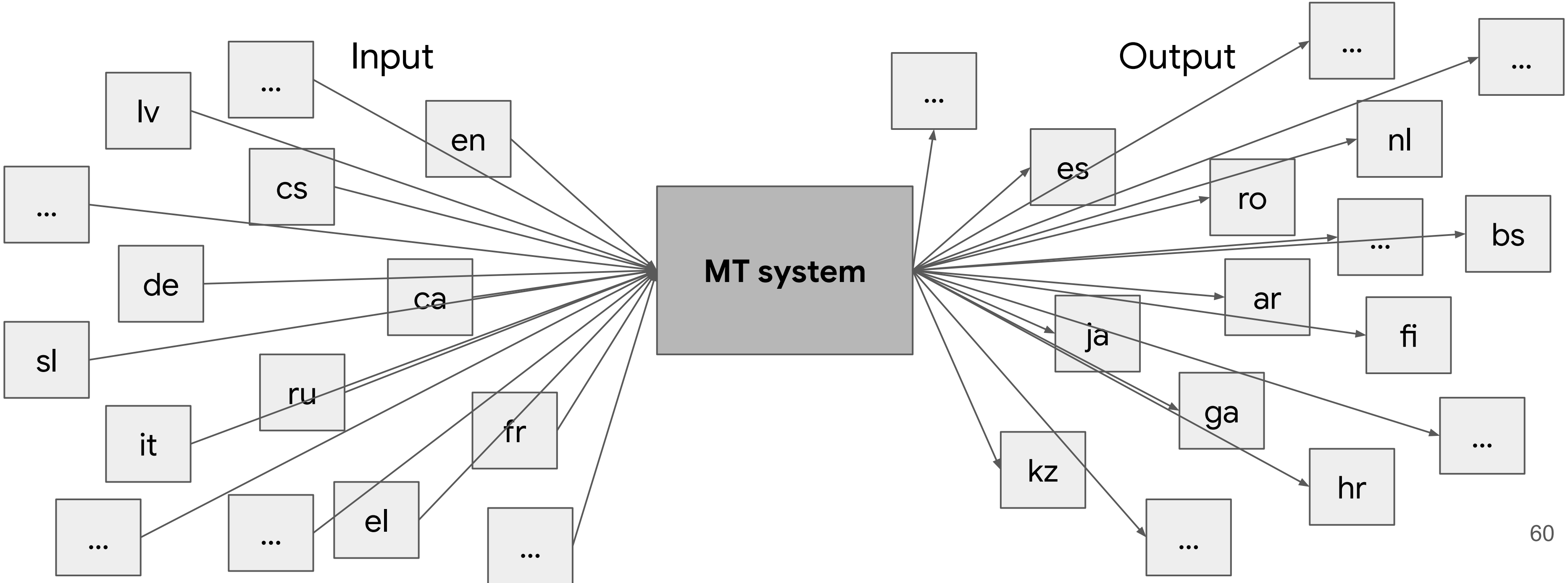
- Word-KD and its enhanced variants
- Best-performing KD methods not applied
- English-centric
- Multi-teacher distillation
- Comparison with other distillation strategies
- Same architecture for teacher and student ⟶

*Can we **improve** performance via KD?*

Methods	WMT'14 En-De	
	BLEU	COMET
Student (Transformer$_{base}$)	$27.42_{\pm 0.01}$	$48.11_{\pm 1.04}$
+ Word-KD (Kim and Rush, 2016)	$28.03_{\pm 0.10}$	$51.59_{\pm 0.23}$
+ Seq-KD (Kim and Rush, 2016)	$28.22_{\pm 0.02}$	$51.23_{\pm 0.15}$
+ Annealing KD (Jafari et al., 2021)	$27.91_{\pm 0.10}$	$51.58_{\pm 0.03}$
+ Selective-KD (Wang et al., 2021)	$28.24_{\pm 0.21}$	$52.15_{\pm 0.42}$
+ TIE-KD (ours)	$\mathbf{28.46}^{*}_{\pm 0.01}$	$\mathbf{52.63}^{*}_{\pm 0.09}$
Teacher (Transformer$_{big}$)	*28.81*	*53.20*

Massively Multilingual MT

What a single MT model to translate from many into many languages (Aharoni et al., 2019).



Multilingual MT

Key Studies

[1] Mohammadshahi et al. (2022)

- **Teacher:** M2M-100 (1.2B)
- **Student**: Deep encoder / shallow decoder (330M)
- **Strategy**: Word-KD + Uniform sub-sampling

Multilingual MT

Key Studies

[1] Mohammadshahi et al. (2022)

- **Teacher:** M2M-100 (1.2B)
- **Student**: Deep encoder / shallow decoder (330M)
- **Strategy**: Word-KD + Uniform sub-sampling

[2] Bapna et al. (2022)

- **Teacher:** 6B
- **Student**: Shallow encoder (330M)
Deep encoder (850M)
- **Strategy**: Seq-KD, Forward translation + Back-translation, Data filtering

Multilingual MT

Key Studies

[3] NLLB Team et al. (2022)

Wikipedia experiment:

- **Teacher:** 1.3B
- **Student**: 500M
- **Strategy**: **Seq-KD** / Word-KD

Multilingual MT

Key Studies

[3] NLLB Team et al. (2022)

MoE experiment:

- **Teacher:** MoE 54B
- **Student**: 1.3B / 615M
- **Strategy**: Word-KD

		eng_Latn-xx				xx-eng_Latn				xx-yy	Avg.
	size	all	high	low	v.low	all	high	low	v.low	all	all
NLLB-200	54B	45.3	54.9	41.9	39.5	56.8	63.5	54.4	54.4	42.7	48.3
dense baseline	1.3B	43.5	52.8	40.1	37.6	54.7	61.8	52.2	51.9	41.0	46.4
dense distilled	1.3B	**44.0**	**53.2**	**40.8**	**38.4**	**55.1**	**61.9**	**52.6**	**52.5**	**41.5**	**46.9**
dense baseline	615M	41.4	50.7	38.1	35.1	52.2	59.7	49.6	49.1	39.3	44.3
dense distilled	615M	**41.8**	**50.9**	**38.5**	**35.8**	**52.3**	59.7	**49.7**	**49.3**	**39.5**	**44.6**

Table 41: **Distillation of NLLB-200.** We report chrF++ scores on FLORES-200 devtest set for the full NLLB-200, dense baselines, and dense distilled models. For eng_Latn-xx and xx-eng_Latn we include all 201 pairs each. For xx-yy we randomly choose 200 directions. We observe that distilled models perform better than dense baseline models trained from scratch without distillation.

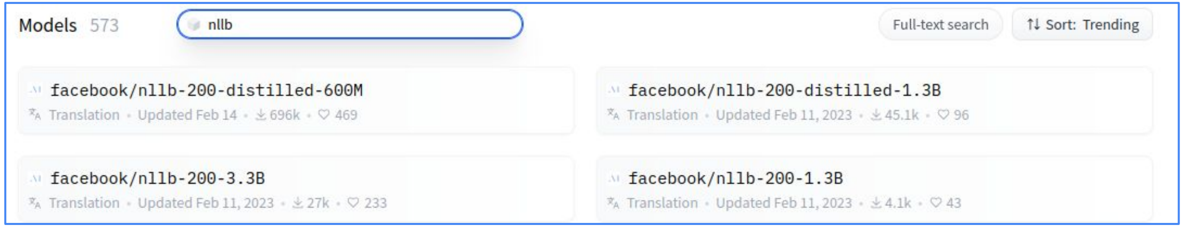
Multilingual MT

Key Studies

[3] NLLB Team et al. (2022)

MoE experiment:

- **Teacher:** MoE 54B
- **Student**: 1.3B / 615
- **Strategy**: Word-KD



		eng_Latn-xx				xx-eng_Latn				xx-yy	Avg.
	size	all	high	low	v.low	all	high	low	v.low	all	all
NLLB-200	54B	45.3	54.9	41.9	39.5	56.8	63.5	54.4	54.4	42.7	48.3
dense baseline	1.3B	43.5	52.8	40.1	37.6	54.7	61.8	52.2	51.9	41.0	46.4
dense distilled	1.3B	**44.0**	**53.2**	**40.8**	**38.4**	**55.1**	**61.9**	**52.6**	**52.5**	**41.5**	**46.9**
dense baseline	615M	41.4	50.7	38.1	35.1	52.2	59.7	49.6	49.1	39.3	44.3
dense distilled	615M	**41.8**	**50.9**	**38.5**	**35.8**	**52.3**	59.7	**49.7**	**49.3**	**39.5**	**44.6**

Table 41: **Distillation of NLLB-200.** We report chrF++ scores on FLORES-200 devtest set for the full NLLB-200, dense baselines, and dense distilled models. For eng_Latn-xx and xx-eng_Latn we include all 201 pairs each. For xx-yy we randomly choose 200 directions. We observe that distilled models perform better than dense baseline models trained from scratch without distillation.

Massively Multilingual MT

Takeaways

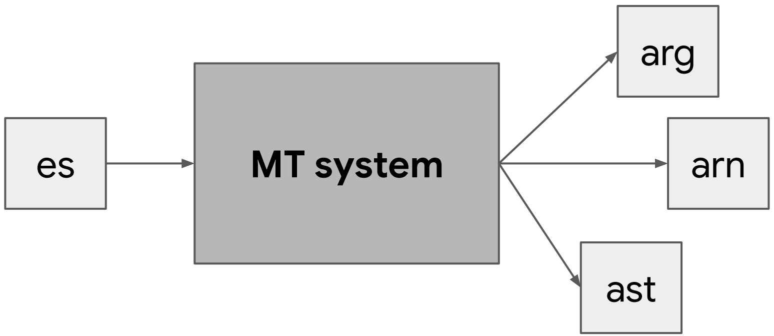
- Seq-KD gives better results but is more expensive
- Many-to-many
- Single teacher distillation
- Deep encoders
- Comparison with teacher performance
- On average 26 times smaller students → *How to best* ***compress*** *knowledge?*

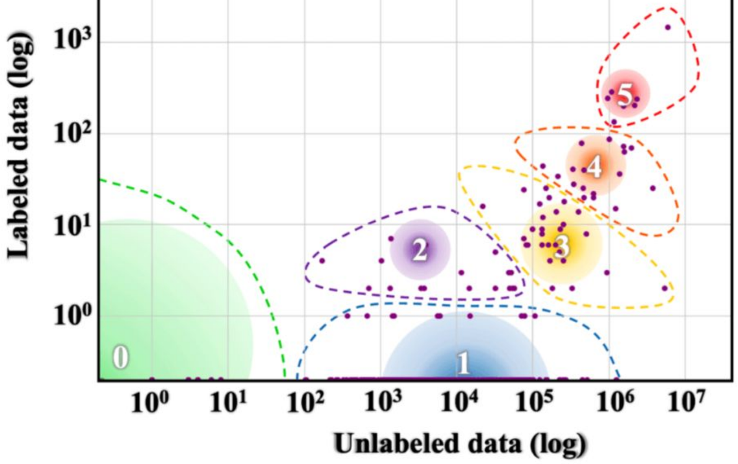
Low-resource MT

What MT that involves languages with limited amount of training data (Haddow et al. 2022).

Input

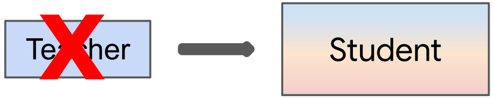
Output





Language Resource Distribution
Joshi et al. (2020)

Low-resource MT

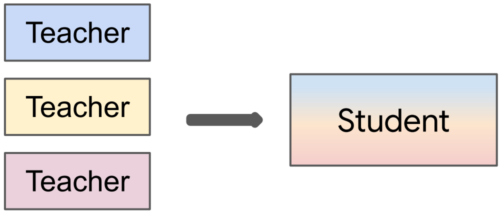


Key studies - *Can we* ***improve*** *performance via KD?*

[1] No teacher available

1. Use of monolingual data
    - Word Similarity Distillation (Zhang et al., 2020)
    - Use an LM to regularize MT outputs (Baziotis et al., 2020)
2. Pivot-based distillation (Chen et al., 2017; He et al., 2019; Ahmed et al., 2024)

Low-resource MT



Key studies - *Can we **improve** performance via KD?*

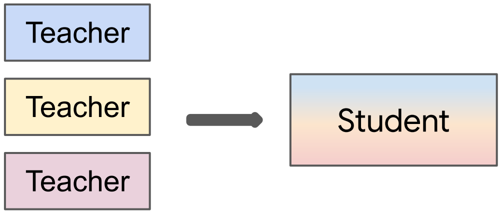
[2] Multi-teacher distillation (building on top of Tan et al., 2019)

1. Adaptive Word-KD (Saleh et al., 2020)

   Access to HRL MT + LRL data

   1. Fine-tune HRL MT with LRL data to train several bilingual teachers
   2. Use the teachers with adaptive KD to train a multilingual student
   3. Dynamically adjust the contribution weight of each teacher

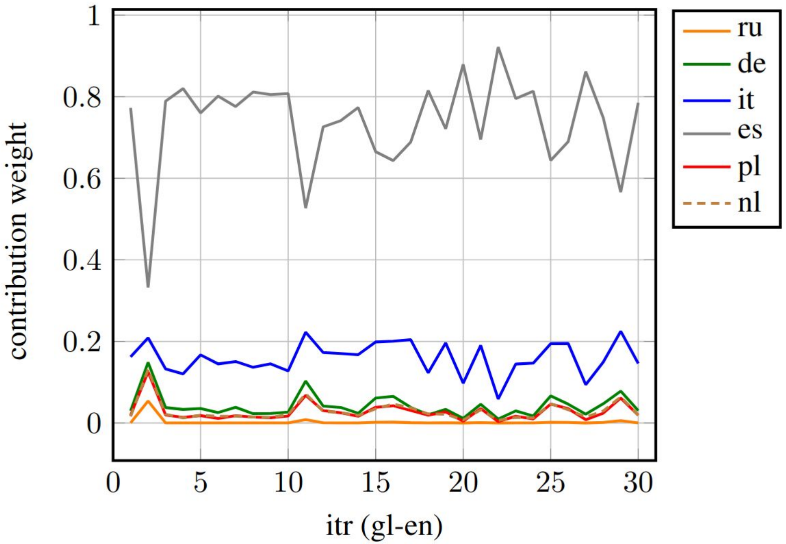
Low-resource MT



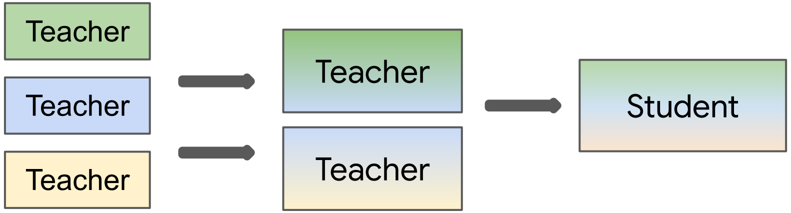
Key studies - *Can we* ***improve*** *performance via KD?*

[2] Multi-teacher distillation (building on top of Tan et al., 2019)

1. Adaptive Word-KD (Saleh et al., 2020)



Low-resource MT



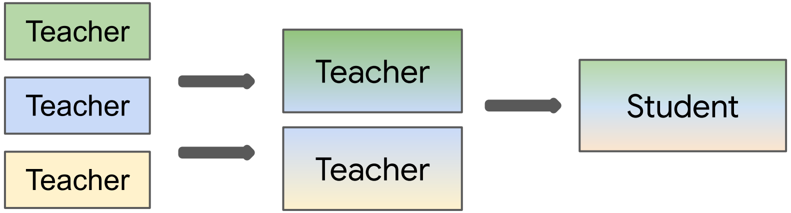
Key studies - *Can we* ***improve*** *performance via KD?*

[2] Multi-teacher distillation (building on top of Tan et al., 2019)

1. Adaptive Word-KD (Saleh et al., 2020)
2. Hierarchical Word-KD (Saleh et al., 2021)

Negative transfer might occur when using multiple teachers

Low-resource MT



Key studies - *Can we **improve** performance via KD?*

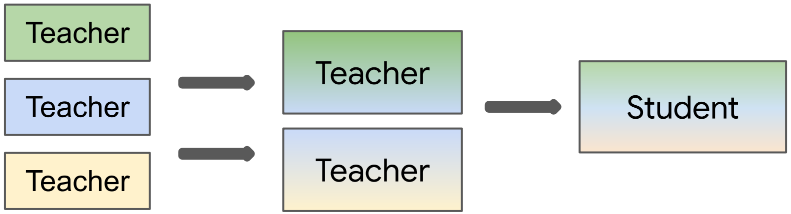
[2] Multi-teacher distillation (building on top of Tan et al., 2019)

1. Adaptive Word-KD (Saleh et al., 2020)
2. Hierarchical Word-KD (Saleh et al., 2021)

Negative transfer might occur when using multiple teachers.

1. Train individual teachers
2. Cluster languages into teacher-assistant models
3. Train super multilingual student

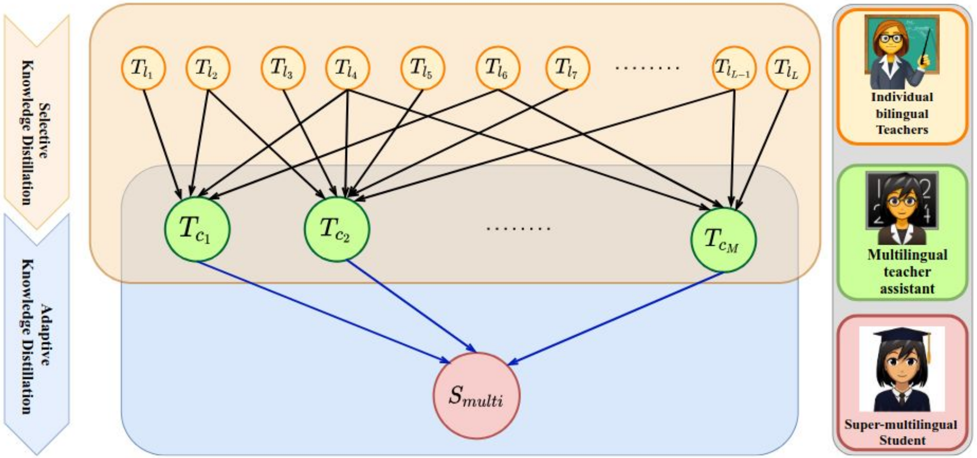
Low-resource MT



Key studies - *Can we **improve** performance via KD?*

[2] Multi-teacher distillation (building on top of Tan et al., 2019)

1. Adaptive Word-KD (Saleh et al., 2020)
2. Hierarchical Word-KD (Saleh et al., 2021)



Low-resource MT

Key studies - *Can we* ***improve*** *performance via KD?*

[3] Pre-trained models and Seq-KD

1. mBART50 (Galiano-Jimémez et al., 2023)
2. NLLB (Song et al., 2023)

Low-resource MT

Key studies - *How to best* ***compress*** *knowledge?*

[1] Model compression

1. Transfer Learning + Seq-KD (Dabre and Fujita., 2020)
2. Priors of Seq-KD vs Quantization (Diddee et al., 2022)
3. Seq-KD Compression of MNMT (Gumma et al., 2023)

Low-resource MT

Key studies - *How to best **compress** knowledge?*

[1] Model compression

1. Transfer Learning + Seq-KD (Dabre and Fujita., 2020)
    - TL: train a model with a HRL and a LRL
    - KD: use the model to create a distilled dataset
2. Priors of Seq-KD vs Quantization (Diddee et al., 2022)
    - Priors: amount of data, student architecture, hyper-parameters
    - Seq-KD gives better results
    - Quantization is more stable
3. Seq-KD Compression of MNMT (Gumma et al., 2023)

Low-resource MT

Key studies - *How to best **compress** knowledge?*

[1] Model compression

3. Seq-KD Compression of MNMT (Gumma et al., 2023)
   - Seq-KD works!

Selective Distillation (Wang et al., 2021)

Lang	OG_base	IT	SLD	W+S LD	BL	GL	GLwD
as	18.4	23.3	19.7	19.8	**20.5**	20.3	**20.5**
bn	28.9	31.8	28.8	28.9	**29.1**	28.3	28.7
gu	30.6	34.1	30.6	31.5	**31.7**	31.3	30.9
hi	34.3	37.5	34.1	34.2	**34.7**	34.4	34.6
kn	25.2	28.7	**26.1**	25.8	25.9	26.0	25.8
ml	27.7	31.4	**28.2**	27.9	**28.2**	27.6	28.0
mr	27.4	31.0	**28.1**	28.0	27.8	27.5	27.8
or	26.3	29.8	26.8	27.0	27.0	**27.1**	26.5
pa	31.0	35.8	31.2	**31.4**	31.3	**31.4**	31.1
ta	25.3	28.4	25.1	25.1	**25.4**	25.2	25.2
te	30.4	33.4	30.4	**30.6**	30.2	**30.6**	30.4
Avg	27.8	31.4	28.1	28.2	28.3	28.2	28.1

Table 3: BLEU scores of base model *distilled* with various distillation techniques. Note that the scores of the *base* model trained on the Original Samanantar data (OG_base) and IndicTrans (IT; *huge*) in the first and second columns are for reference. The best scores of distilled models are bolded.

Low-resource MT

Takeaways

- Studies with different goals
- English-centric translation
- Promising avenues:
    - Seq-KD
    - Pre-trained models
    - LLMs? (Enis and Hopkins, 2024)

KD4MT @ Helsinki-NLP

Tools: OpusDistillery

- OpusDistillery is an end-to-end pipeline

    systematic

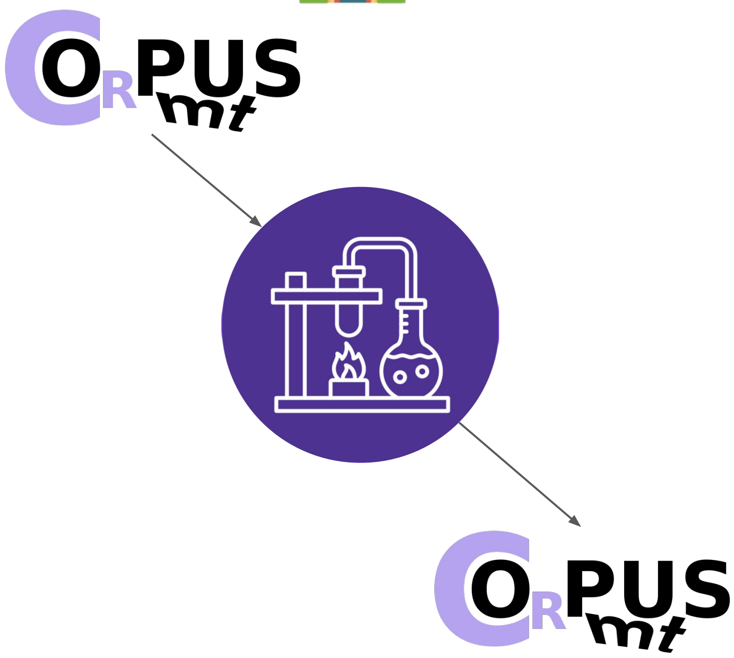
    multilingual

    distillation of

    OPUS-MT

    Models

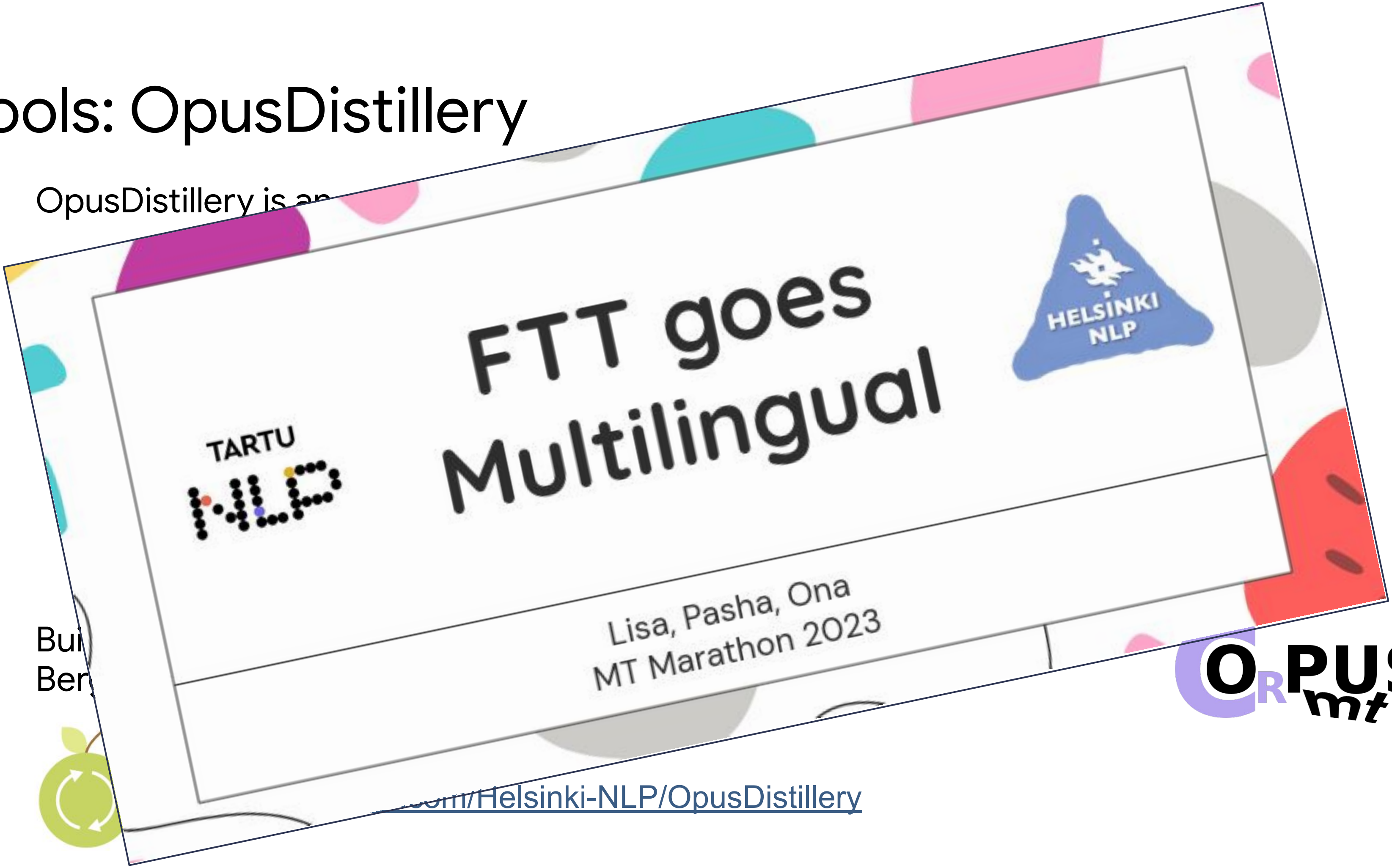
- Built on top of open-source tools from the Bergamot project



https://github.com/Helsinki-NLP/OpusDistillery

Tools: OpusDistillery

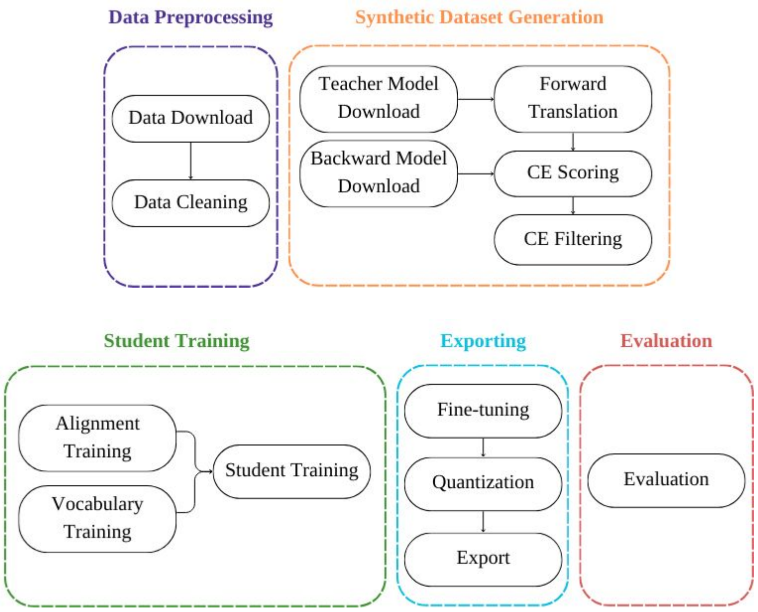
- OpusDistillery is an
- Bui
  Ber



[...com/Helsinki-NLP/OpusDistillery](https://github.com/Helsinki-NLP/OpusDistillery)

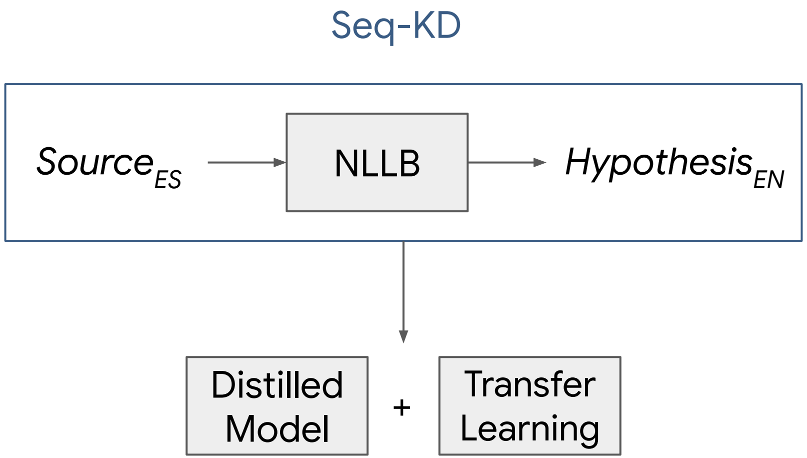
Tools: OpusDistillery

```yaml
experiment:
  dirname: base-multi
  name: eng-zle
  langpairs:
    - en-uk
    - en-ru
    - en-be

  #URL to the OPUS-MT model to use as the teacher
  opusmt-teacher:
  - "https://object.pouta.csc.fi/Tatoeba-MT-models/
      eng-sla/opus2m-2020-08-01.zip"

  #URL to the OPUS-MT model to use as the backward model
  opusmt-backward: "https://object.pouta.csc.fi/Tatoeba-MT-models/
      sla-eng/opus4m-2020-08-12.zip"

  # Specify if the teacher, the student and the backward models
  # are many2one to deal with language tags
  one2many-teacher: True
  one2many-student: True
  one2many-backward: false

  teacher-ensemble: 1

  parallel-max-sentences: 10000000
  split-length: 1000000

  best-model: perplexity
  spm-sample-size: 1000000

datasets:
  train:
    - tc_Tatoeba-Challenge-v2023-09-26
  devtest:
    - flores_dev
  test:
    - flores_devtest
```



Low-resource MT

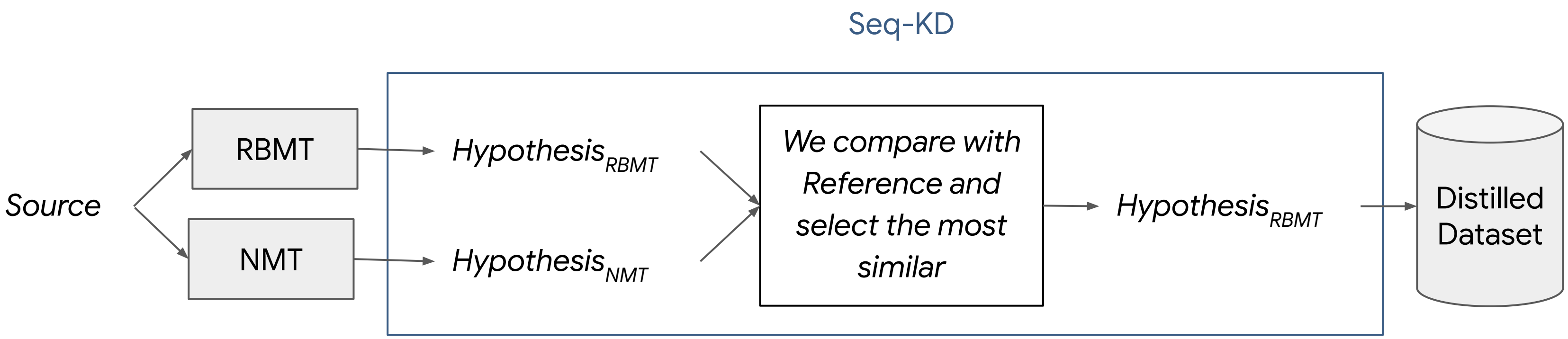
1. **Americas NLP 2023 Shared Task** on Machine Translation into Indigenous Languages
    - Spanish > 11 indigenous languages of the Americas
    - We use **Seq-KD** to reduce the size of a large model (NLLB) and enable efficient fine-tuning



https://turing.iimas.unam.mx/americasnlp/2023_st.html

Low-resource MT

2. WMT24 **Translation into Low-Resource Languages of Spain** Shared Task
    - Spanish > Aragonese, Asturian, Occitan (Gascon Variant)
    - We use Seq-KD to benefit from both the RBMT and the NMT systems



Low-resource MT

2. WMT24 **Translation into Low-Resource Languages of Spain** Shared Task
   - Spanish > Aragonese, Asturian, Occitan (Gascon Variant)
   - We use Seq-KD to benefit from both the RBMT and the NMT systems

Apertium

#	Method	BLEU / ChrF			Params (M)	Speed (s)
		arg	arn	ast		
1	Fine-tuning Data Sampling Ensembling	51.5 / 75.6	22.1 / 45.1	18.2 / 51.6	222.9	852.22
2	Distillation RBMT+NMT Ensembling	50.6 / 75.4	22.4 / 45.7	18.0 / 51.6	65.7	361.33
3	Distillation RBMT+NMT	49.1 / 75.4	21.6 / 45.0	17.9 / 51.4	20.4	4.06
Best	–	63.0 / 80.3	30.1 / 50.1	23.2 / 55.2	–	–

Table 5: Summary of our submissions. BLEU refers to the score obtained by the best ensemble on the development set; Speed refers to the averaged decoding speed for submission across language pairs on one single AMD MI250x GPU. In addition, we provide the best competitor scores for each target language.

Low-resource MT

2. WMT24 **Translation into Low-Resource Languages of Spain** Shared Task
   - Spanish > Aragonese, Asturian, Occitan (Gascon Variant)
   - We use Seq-KD to benefit from both the RBMT and the NMT systems

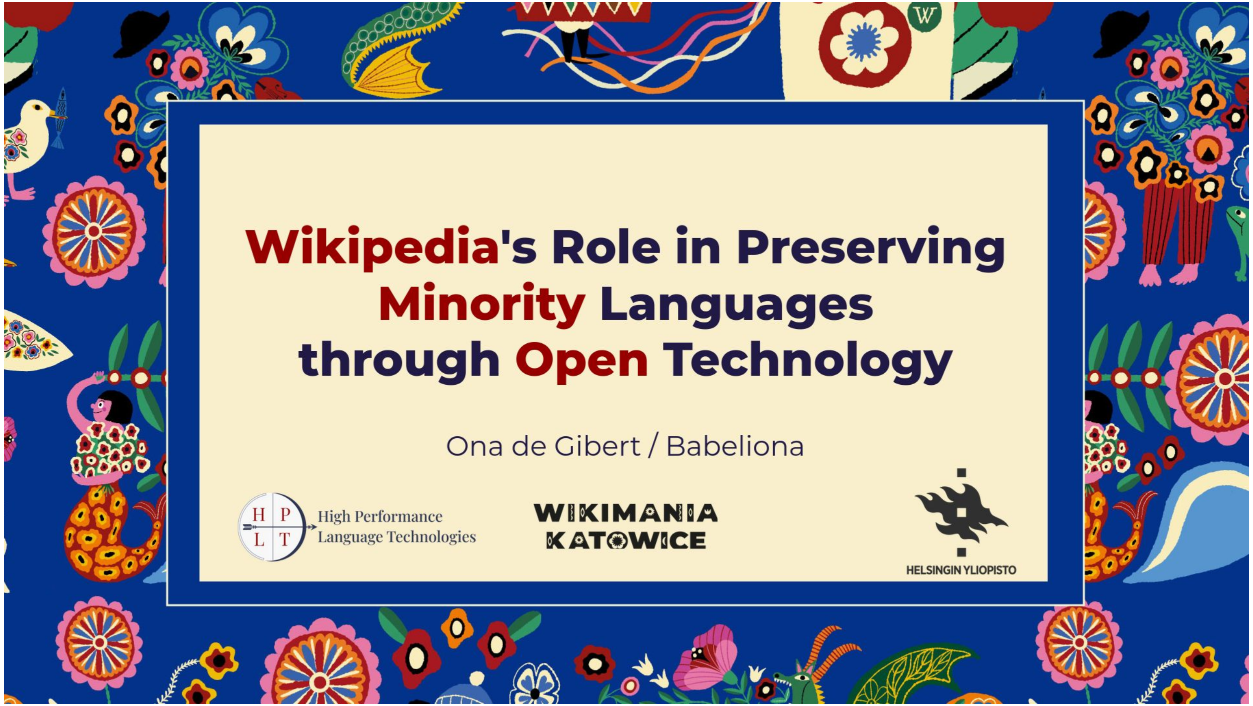
Apertium

#	Method	BLEU / ChrF arg	BLEU / ChrF arn	BLEU / ChrF ast	Params (M)	Speed (s)
1	Fine-tuning Data Sampling Ensembling	51.5 / 75.6	22.1 / 45.1	18.2 / 51.6	222.9	852.22
2	Distillation RBMT+NMT Ensembling	50.6 / 75.4	22.4 / 45.7	18.0 / 51.6	65.7	361.33
3	Distillation RBMT+NMT	49.1 / 75.4	21.6 / 45.0	17.9 / 51.4	20.4	4.06
Best	–	63.0 / 80.3	30.1 / 50.1	23.2 / 55.2	–	–

Table 5: Summary of our submissions. BLEU refers to the score obtained by the best ensemble on the development set; Speed refers to the averaged decoding speed for submission across language pairs on one single AMD MI250x GPU. In addition, we provide the best competitor scores for each target language.

https://www2.statmt.org/wmt24/romance-task.html

Unexpected Bonus: MT at Wikipedia!



Machine Translation at Wikipedia

- **Apertium** - 34 languages
- **MinT** - 236 languages

- **Elia** - 6 languages
- **Google Translate** - 135 languages
- **LingoCloud** - 5 languages
- **Yandex** - 99 languages

WIKIMANIA KATOWICE

Open Machine Translation at Wikipedia

- **Apertium** - 34 languages
- **MinT** - 236 languages

- **Elia** - 6 languages
- **Google Translate** - 135 languages
- **LingoCloud** - 5 languages
- **Yandex** - 99 languages

WIKIMANIA KATOWICE

Open Machine Translation at Wikipedia

- **MinT**
  - self hosted Neural Machine Translation service by Wikipedia
  - more than 70 languages not supported by other services!
  - several open-source initiatives
    - NLLB
    - SoftCatala
    - IndicTrans2
    - OpusMT
    - MADLAD-400

WIKIMANIA KATOWICE

Open **Machine Translation at Wikipedia**

- Wikimedia does not run any proprietary software
- MinT translation services uses **quantized models**
- Two issues:
  - Cost
  - Propietary drivers

Open fast MT models on CPU

HPLT High Performance Language Technologies

HELSINGIN YLIOPISTO

WIKIMANIA KATOWICE

Laxström, N., & Thottingal, S. (2023). Machine Translation at Wikipedia. Workshop on Open Community-Driven Machine Translation, EAMT 2023, Tampere.

Future of KD4MT

What are the research gaps?

Future of KD4MT

- What exactly happens during KD? Gender bias, Uncertainty, Robustness...

Future of KD4MT

- What exactly happens during KD? Gender bias, Uncertainty, Robustness…
- What is the optimal teacher?
    - Capacity gap
        - if we gradually increase the size of the teacher, the performance of the student improves for a while and then it starts to drop (Mirzadeh et al., 2019)
        - Increasing the size of the teacher usually boosts its performance, but does not necessarily lead to a better teacher for the student

Future of KD4MT

- What exactly happens during KD? Gender bias, Uncertainty, Robustness...
- What is the optimal teacher?
    - Capacity gap
        - if we gradually increase the size of the teacher, the performance of the student improves for a while and then it starts to drop (Mirzadeh et al., 2019)
        - Increasing the size of the teacher usually boosts its performance, but does not necessarily lead to a better teacher for the student
- What is the optimal student architecture?

Future of KD4MT

- What exactly happens during KD? Gender bias, Uncertainty, Robustness...
- What is the optimal teacher?
  - Capacity gap
    - if we gradually increase the size of the teacher, the performance of the student improves for a while and then it starts to drop (Mirzadeh et al., 2019)
    - Increasing the size of the teacher usually boosts its performance, but does not necessarily lead to a better teacher for the student
- What is the optimal student architecture?
- Do the current KD methods generalize in multilingual setups?

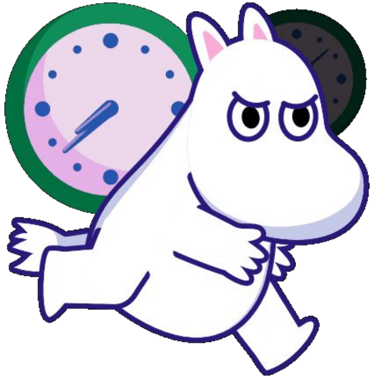
Future of KD4MT

- What exactly happens during KD? Gender bias, Uncertainty, Robustness...
- What is the optimal teacher?
  - Capacity gap
    - if we gradually increase the size of the teacher, the performance of the student improves for a while and then it starts to drop (Mirzadeh et al., 2019)
    - Increasing the size of the teacher usually boosts its performance, but does not necessarily lead to a better teacher for the student
- What is the optimal student architecture?
- Do the current KD methods generalize in multilingual setups?
- What about non english-centric setups?

Future of KD4MT

- What exactly happens during KD? Gender bias, Uncertainty, Robustness...
- What is the optimal teacher?
    - Capacity gap
        - if we gradually increase the size of the teacher, the performance of the student improves for a while and then it starts to drop (Mirzadeh et al., 2019)
        - Increasing the size of the teacher usually boosts its performance, but does not necessarily lead to a better teacher for the student
- What is the optimal student architecture?
- Do the current KD methods generalize in multilingual setups?
- What about non english-centric setups?
- Can we integrate LLMs in the distillation process for MT?

Thanks for listening!
Questions?



References

Aji, A. F., & Heafield, K. (2020). Fully Synthetic Data Improves Neural Machine Translation with Knowledge Distillation.

Gumma, V., Dabre, R., & Kumar, P. (2023). An Empirical Study of Leveraging Knowledge Distillation for Compressing Multilingual Neural Machine Translation Models. Proceedings of the 24th Annual Conference of the European Association for Machine Translation.

Jafari, A., Rezagholizadeh, M., Sharma, P., & Ghodsi, A. (2021). Annealing Knowledge Distillation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2493–2504, Online. Association for Computational Linguistics.

Kim, Y., & Rush, A. M. (2016). Sequence-Level Knowledge Distillation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.

Wang, F., Yan, J., Meng, F., & Zhou, J. (2021). Selective Knowledge Distillation for Neural Machine Translation. 6456-6466. 10.18653/v1/2021.acl-long.504.

Wu, Y., Passban, P., & Rezagholizadeh, M. (2020). Why Skip If You Can Combine: A Simple Knowledge Distillation Technique for Intermediate Layers. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

Yang, Z., Sun, R., & Wan, X. (2022). Nearest Neighbor Knowledge Distillation for Neural Machine Translation. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

Zhang, S., Liang, Y., Wang, S., Chen, Y., Han, W., Liu, J., & Xu, J. (2023). Towards Understanding and Improving Knowledge Distillation for Neural Machine Translation. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.

References

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. ACM Computing Surveys (CSUR), 53(5):1–38.

Heejin Do and Gary Geunbae Lee. 2023. Target-oriented knowledge distillation with language-family-based grouping for multilingual nmt. ACM Trans. Asian Low-Resour. Lang. Inf. Process., 22(2).

Yichong Huang, Xiaocheng Feng, Xinwei Geng, and Bing Qin. 2022. Unifying the convergences in multilingual neural machine translation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6822–6835, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. arXiv preprint arXiv:1902.10461.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Bérard, Caroline Brun, James Henderson, and Laurent Besacier. 2022a. Small-100: Introducing shallow multilingual machine translation model for low-resource languages. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8348–8359.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. arXiv preprint arXiv:2205.03983.

References

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, Online. Association for Computational Linguistics.

Xinlu Zhang, Xiao Li, Yating Yang, and Rui Dong. 2020. Improving low-resource neural machine translation with teacher-free knowledge distillation. IEEE Access, 8:206638–206645

Tianyu He, Jiale Chen, Xu Tan, and Tao Qin. 2019. Language graph distillation for low-resource machine translation. arXiv e-prints, pages arXiv–1908.

Sangchul Hahn and Heeyoul Choi. 2019. Self-knowledge distillation in natural language processing. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pages 423–430, Varna, Bulgaria. INCOMA Ltd

Fahimeh Saleh, Wray Buntine, and Gholamreza Haffari. 2020. Collective wisdom: Improving low-resource neural machine translation using adaptive knowledge distillation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3413–3421, Barcelona, Spain (Online). International Committee on Computational Linguistics

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3525–3535, Online. Association for Computational Linguistics.

References

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language Model Prior for Low-Resource Neural Machine Translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7622–7634, Online. Association for Computational Linguistics.

Khalid Ahmed and Jan Buys. 2024. Neural Machine Translation between Low-Resource Languages with Synthetic Pivoting. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 12144–12158, Torino, Italia. ELRA and ICCL.

Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. 2017. A teacher-student framework for zero-resource neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1925–1935

Fahimeh Saleh, Wray Buntine, Gholamreza Haffari, and Lan Du. 2021. Multilingual neural machine translation: Can linguistic hierarchies help? In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1313–1330, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yewei Song, Saad Ezzini, Jacques Klein, Tegawende Bissyande, Clément Lefebvre, and Anne Goujon. 2023. Letz translate: Low-resource machine translation for luxembourgish. In 2023 5th International Conference on Natural Language Processing (ICNLP), pages 165–170. IEEE.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, Víctor M Sánchez-Cartagena, and Juan Antonio Pérez-Ortiz. 2023. Exploiting large pre-trained models for low-resource neural machine translation. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 59–68.

Raj Dabre and Atsushi Fujita. 2020. Combining sequence distillation and transfer learning for efficient low-resource neural machine translation models. In Proceedings of the Fifth Conference on Machine Translation, pages 492–502, Online. Association for Computational Linguistics.

Harshita Diddee, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu, and Kalika Bali. 2022. Too brittle to touch: Comparing the stability of quantization and distillation towards develop- ing low-resource MT models. In Proceedings of the Seventh Conference on Machine Transla- tion (WMT), pages 870–885, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

References

Varun Gumma, Raj Dabre, and Pratyush Kumar. 2023. An empirical study of leveraging knowledge distillation for compressing multilingual neural machine translation models. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 103–114.

Maxim Enis and Mark Hopkins. "From LLM to NMT: Advancing Low-Resource Machine Translation with Claude." arXiv preprint arXiv:2404.13813 (2024).