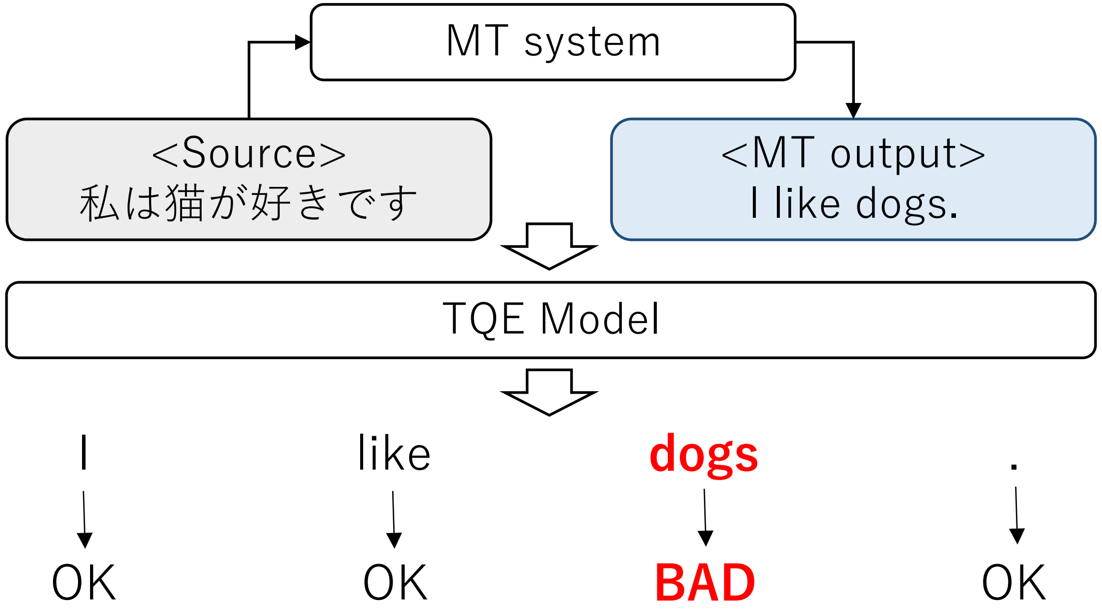
Word-level Translation Quality Estimation Based on Optimal Transport

Yuto Kuroda[1] Atsushi Fujita[2] Tomoyuki Kajiwara[1]
[1] Ehime University, Japan [2] NICT, Japan

AMTA 2024

Introduction

- Translation Quality Estimation (TQE)
  - The task of predicting quality labels or scores for the given translation
  - Sentence-level:
    - Help users determine whether to use an MT output as it is or after post-editing.
  - **Word-level (this work)**:
    - Better guide post-editors in the translation production process, i.e., spotting words that need a revision.



Previous Work [Liu+ 2017; Lee 2020]

- Most work follows a three-step training approach

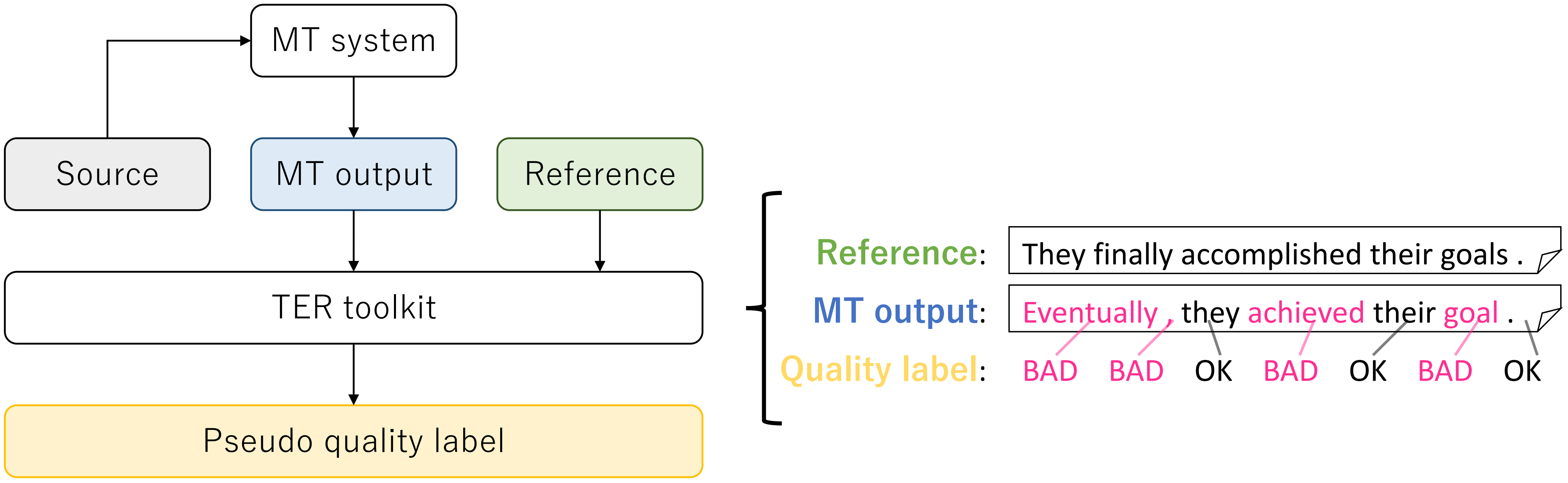
Step	Data Type	Quality Label	Quantity
Step 0. (Encoder) Pre-training	Monolingual/parallel data	n/a	Very large
Step 1. Pre-training for TQE	TQE data (src, mt, label)	Pseudo	Large
Step 2. Fine-tuning	TQE data (src, mt, label)	Manually determined	Small

- Step 1 plays an important role
    - To overcome the data sparseness issue in Step 2
    - Especially for zero-shot translation directions

[Liu+ 2017] Lemao Liu et al. Translation Quality Estimation Using Only Bilingual Corpora. IEEE/ACM TASLP, 2017.
[Lee 2020] Dongjun Lee. Two-Phase Cross-Lingual Language Model Fine-Tuning for Machine Translation Quality Estimation. In Proc. of WMT, 2020.

Previous Work

- Bilingual parallel corpus + MT system + TER toolkit
  → Pseudo-quality label



Problem of Previous Work

- Surface-level differences between independent translations do not necessarily indicate errors.
  - e.g., Synonymous expressions
  - e.g., Interchangeable word orderings

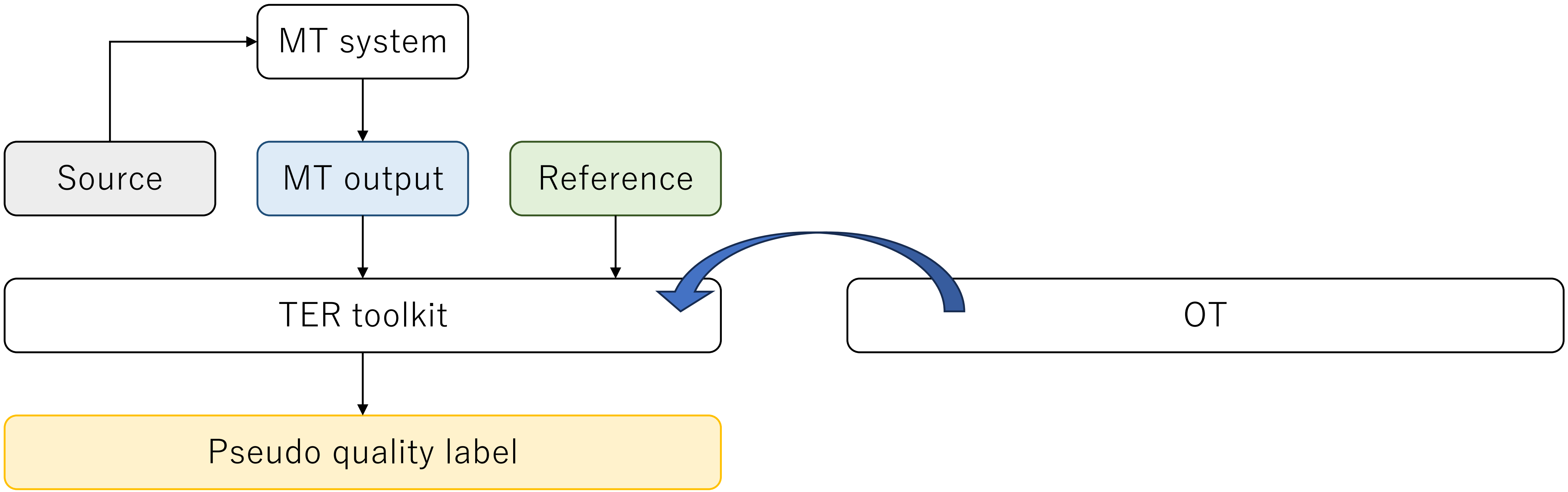
Reference: They finally accomplished their goals .

MT output: Eventually , they achieved their goal .

Quality label: BAD BAD OK BAD OK BAD OK

Proposed Method (Overview)

- Determine pseudo-quality labels using Optimal Transport (OT)
  - inspired by its application to monolingual word alignment [Arase+, 2023]

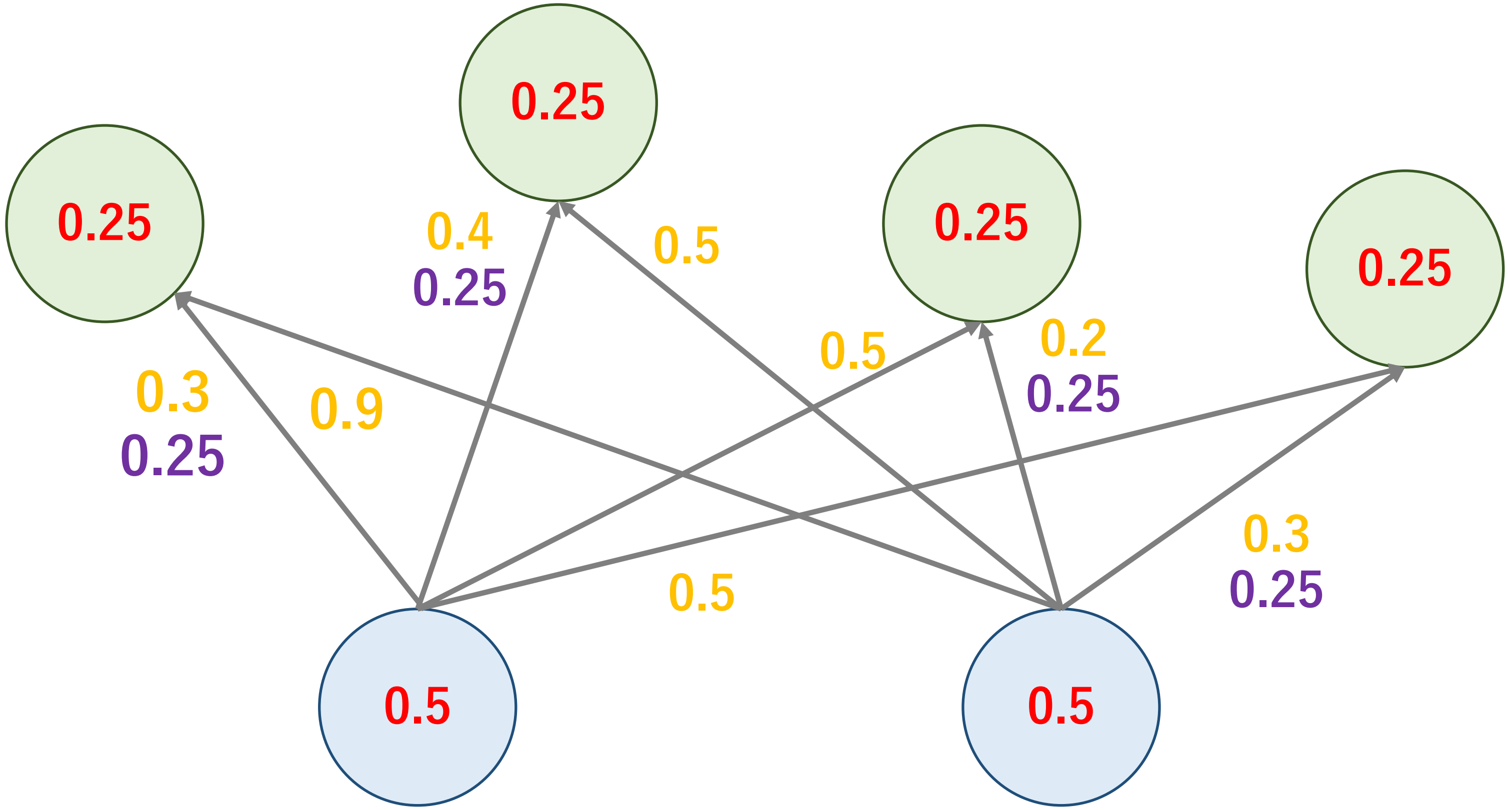


[Arase+ 2023] Yuki Arase et al. Unbalanced Optimal Transport for Unbalanced Word Alignment. In Proc. of ACL, 2023.

Determining Pseudo-Quality Labels Using Optimal Transport

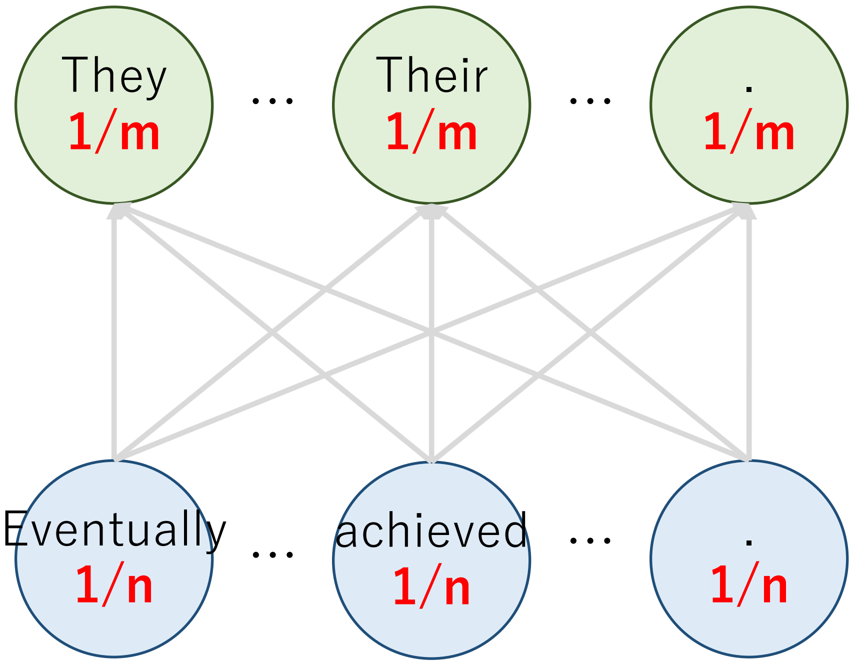
Proposed Method (Basics of OT)

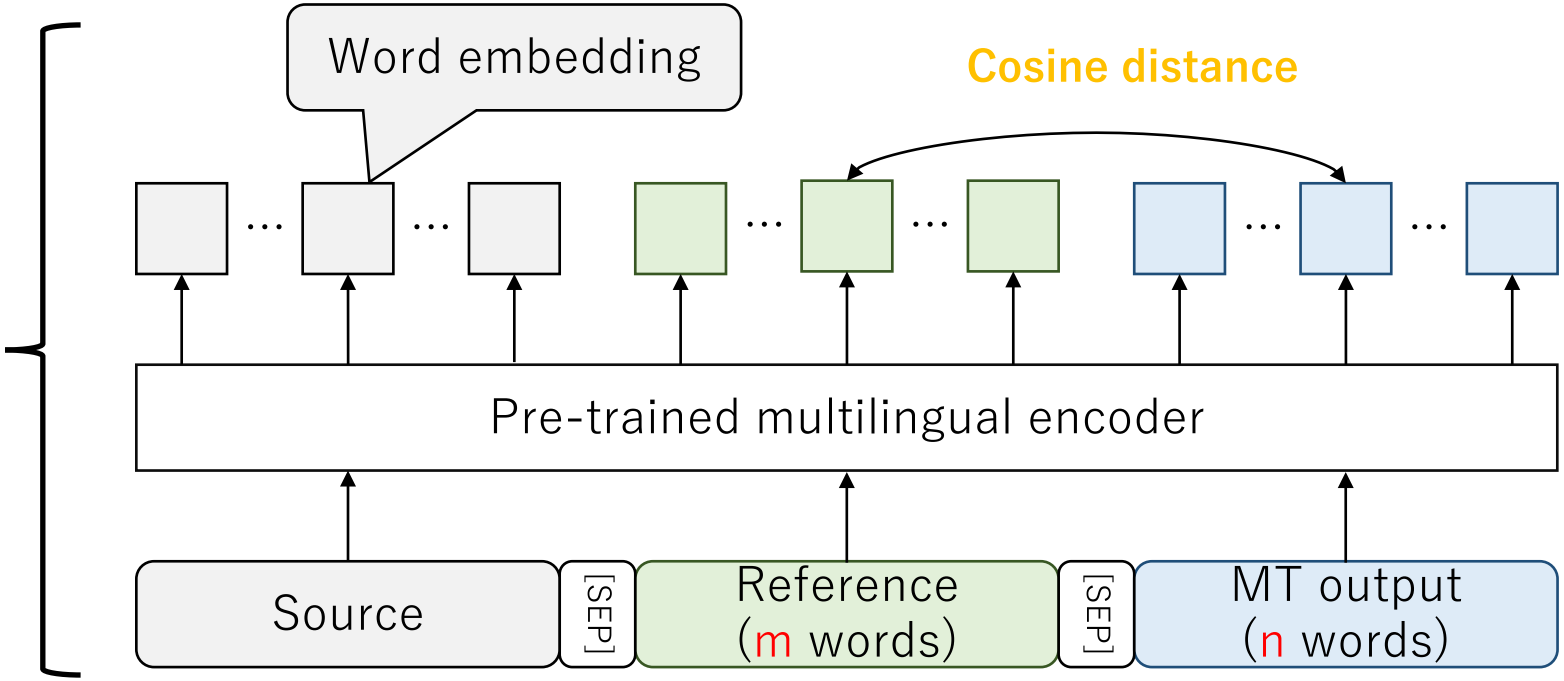
- OT is an algorithm that identifies the optimal way of converting one distribution into another.
  - Input: **Mass of each word (distribution)** and **Cost for transportation**
  - Output: **Optimal transport matrix**



Proposed Method (Components)

- **Mass (weight of tokens)**: uniform distribution
- **Cost**: cosine distance between contextual word embeddings





Proposed Method (Formulation of OT)

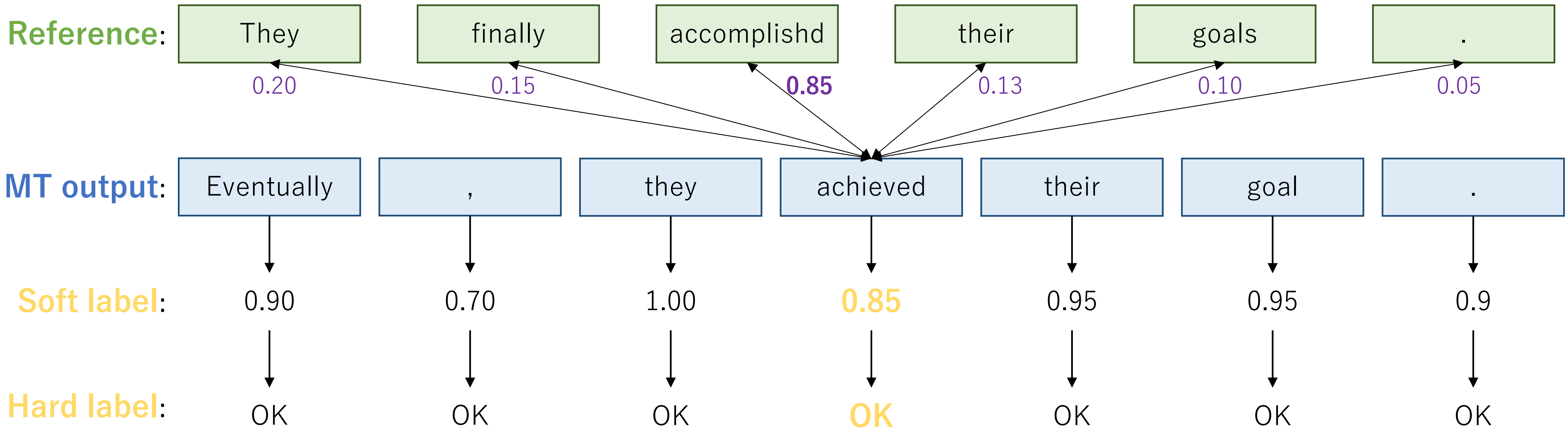
- **Cost Matrix (C)**: cost for all the pairs of words
- **OT Matrix (P)**: minimizes the total cost for transportation.

$$P = \operatorname*{argmin}_{P' \in U} \left(\sum_{i,j} C_{i,j} {P'}_{i,j} - \xi H(P') \right)$$

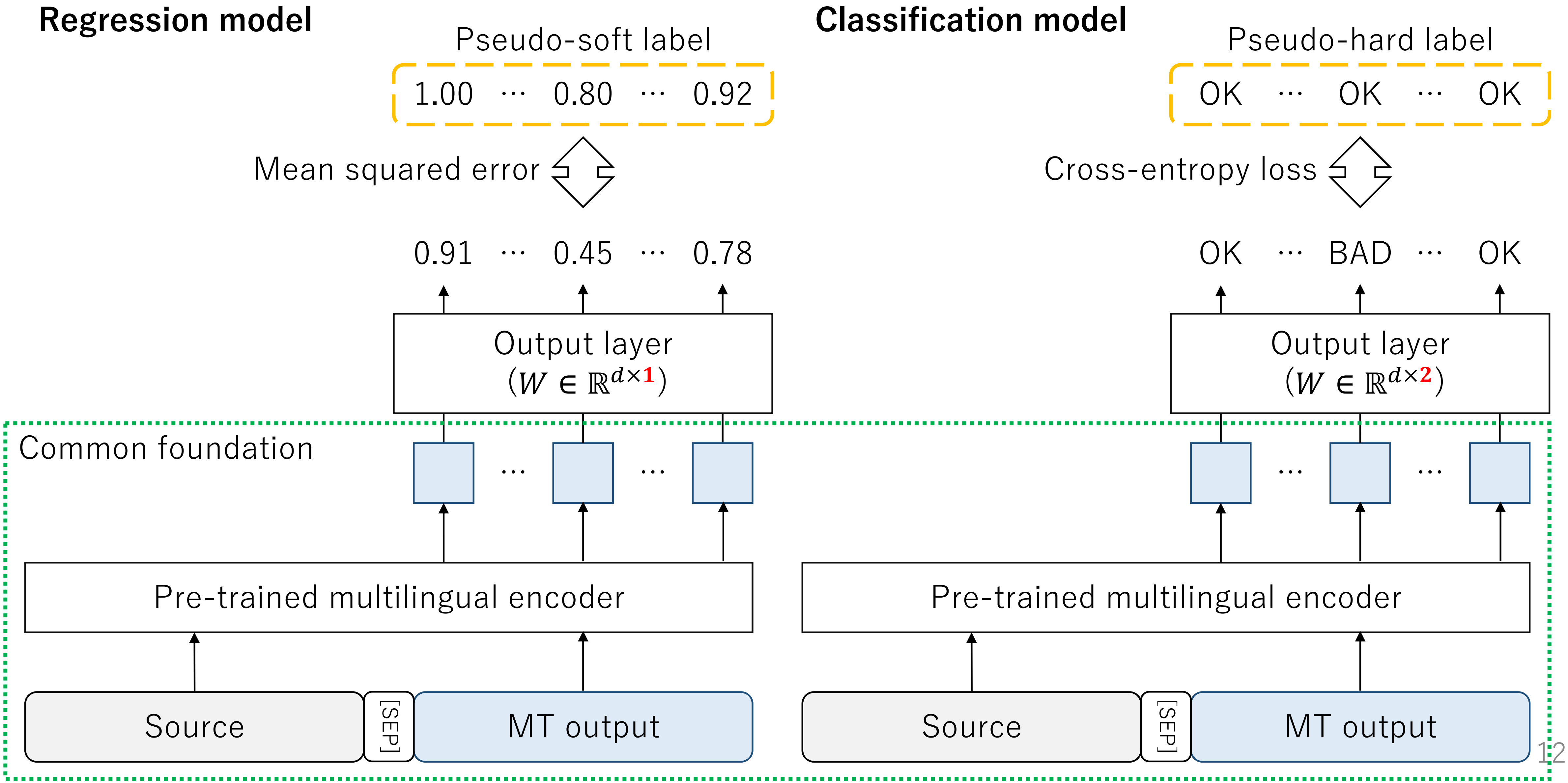
  - $P_{i,j}$: amount of mass to be transferred between each pair of words
  - U : a set of candidate matrices that satisfy several conditions
    - We adopt Partial OT [Figalli 2010; Caffarelli and McCann, 2010]
    - $P 1_n \leq a$: outflow from each word in the **MT output** must be up to 1/n
    - $P^T 1_m \leq b$: inflow into each word in the **Reference** must be up to 1/m
    - $1_n^T P^T 1_m = \lambda_m$: total transportation is bounded to $\lambda_m \in (0, 1]$
  - $H(P')$: entropy-based regularizer (with a weight ξ) [Arase+ 2023]

Proposed Method (Determining Pseudo-Labels)

- **Optimal transport matrix** → **pseudo quality label**
  - Soft label: Maximum amount of mass transferred from the word in the **MT output** text to a word in the **Reference**
  - Hard label: ”OK” or “BAD” determined by thresholding soft label



Two Conventional Architectures of TQE Models



Experiments

Setting

- Dataset
  - Test: MLQE-PE [Fomicheva+ 2022]
    - WMT20 [Specia+ 2020]
    - WMT21 [Specia+ 2021]
    - This talk shows only results for WMT21
  - Training:
    - MLQE-PE Training data
    - Synthetic TQE data
      - Parallel data for WMT21 TQE Task 2
      - Hyper-parameters for OT were optimized on MLQE-PE Dev data
      - Much larger than MLQE-PE Train data

	Language pair	Synthetic data	MLPE-QE WMT21 Train	MLPE-QE WMT21 Dev	MLPE-QE WMT21 Test
Non-zero-shot translation direction	En→De	22,701,552	7,000	1,000	1,000
	En→Zh	16,201,271	7,000	1,000	1,000
	Ro→En	3,027,243	7,000	1,000	1,000
	Et→En	855,680	7,000	1,000	1,000
	Ne→En	166,893	7,000	1,000	1,000
	Si→En	570,770	7,000	1,000	1,000
zero-shot translation direction	En→Cs	—	—	—	1,000
	En→Ja	—	—	—	1,000
	Km→En	—	—	—	990
	Ps→En	—	—	—	1,000
	Ru→En	—	—	—	1,000

[Fomicheva+ 2022] Marina Fomicheva et al. MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset. In Proc. of LREC, 2022.
[Specia+ 2020] Lucia Specia et al. Findings of the WMT 2020 Shared Task on Quality Estimation. In Proc. of WMT, 2020.
[Specia+ 2021] Lucia Specia et al. Findings of the WMT 2021 Shared Task on Quality Estimation. In Proc. of WMT, 2021.

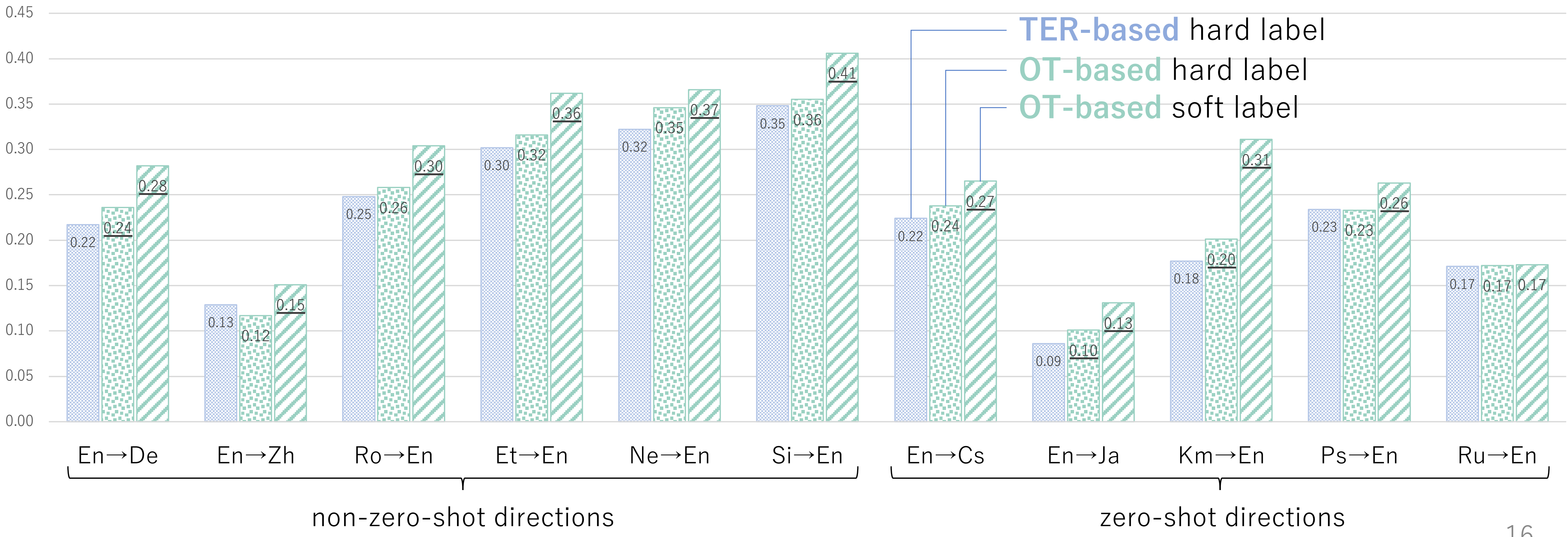
Setting (contd.)

- TQE Models
  - Step 0. (Encoder) Pre-trained Model: InfoXLM$_{\text{Large}}$ [Chi+ 2021]
  - Pseudo-supervised models: Do Step 1 only
    - Baseline: **TER-based** hard labels
    - Proposed: **OT-based** hard labels
    - Proposed: **OT-based** soft labels
  - Fine-tuned models: Do Step 2 (Steps 1&2 or only Step 2)
    - Baseline (only Step 2): Step 2 with MLQE-PE
    - Baseline (Steps 1&2): Step 1 with **TER-based** hard labels + Step 2 with MLQE-PE
    - Proposed (Steps 1&2): Step 1 with **OT-based** hard labels + Step 2 with MLQE-PE
    - Proposed (Steps 1&2): Step 1 with **OT-based** soft labels + Step 2 with MLQE-PE
- Evaluation Metric:
  - Matthews correlation coefficient (MCC) [Matthews 1975]

[Chi+ 2021] Zewen Chi et al. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In Proc. of NAACL, 2021.
[Matthews 1975] B.W. Matthews. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. Biochimica et Biophysica Acta (BBA), 1975.

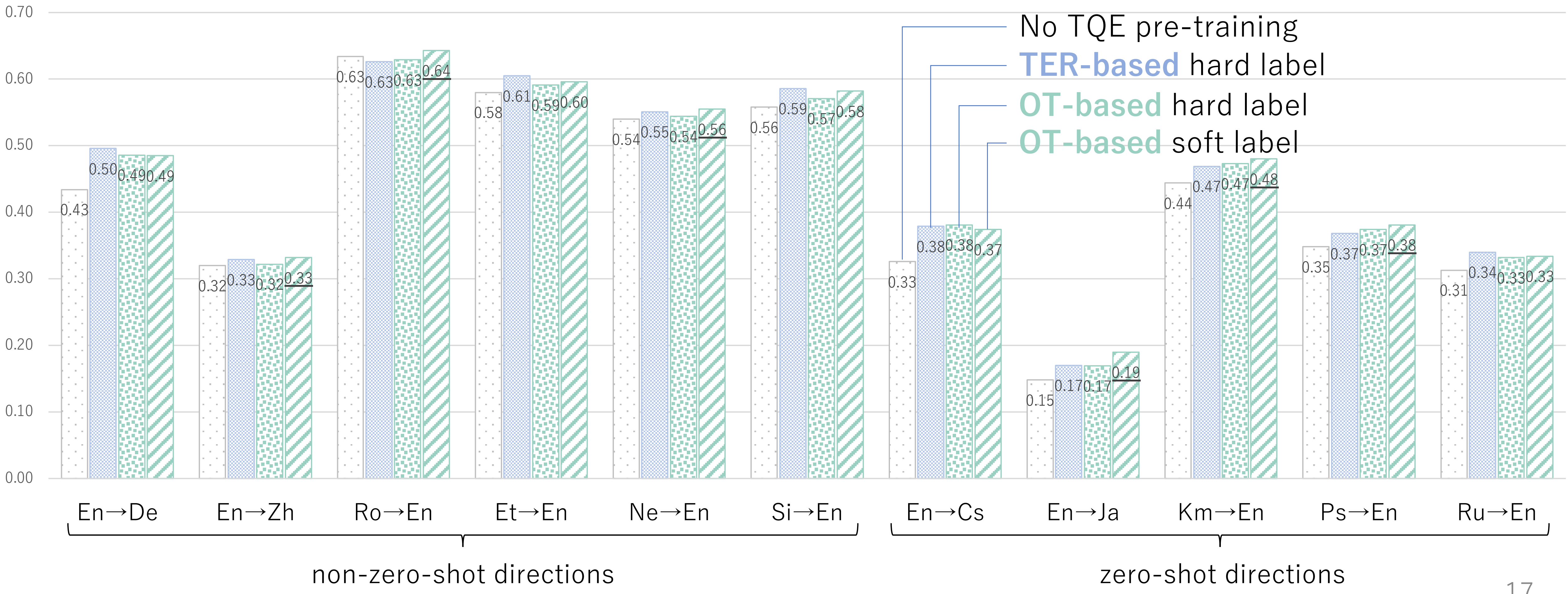
Results (Pseudo-supervised Models)

- The model trained on **OT-based** soft labels outperformed the ones trained on either **TER-based** or **OT-based** hard labels
  - Statistically significant gains over the **TER-based** model (except Ru→En)



Results (Fine-tuned Models)

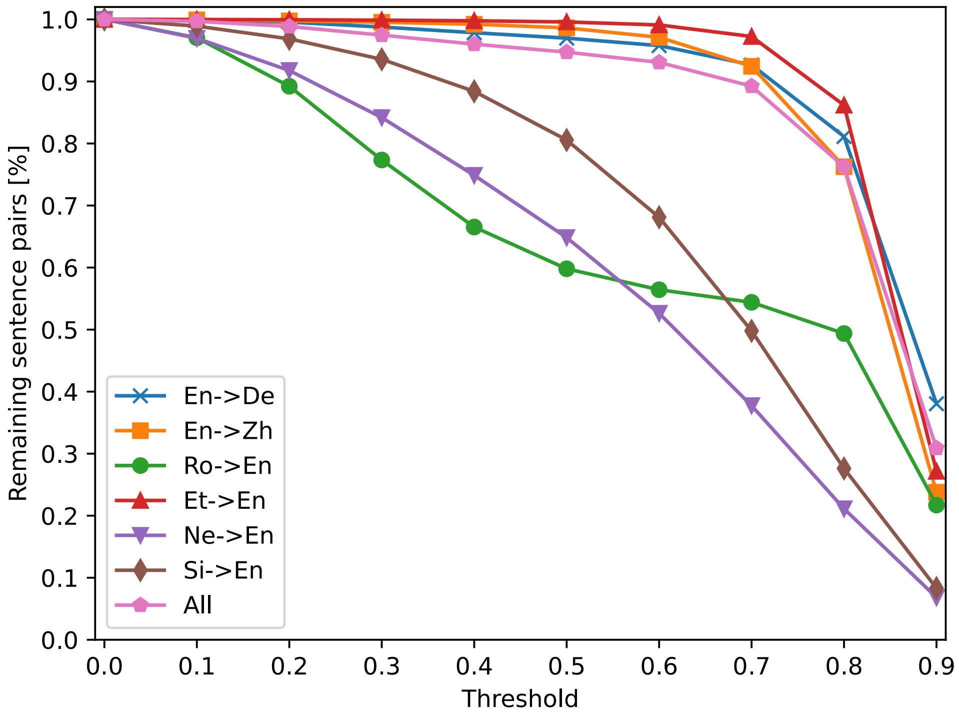
- The model pre-trained on **OT-based** soft labels achieved higher MCC than **TER-based** model for 6 out of the 11 test sets



Analyses

Impact of Synthetic Data Quality

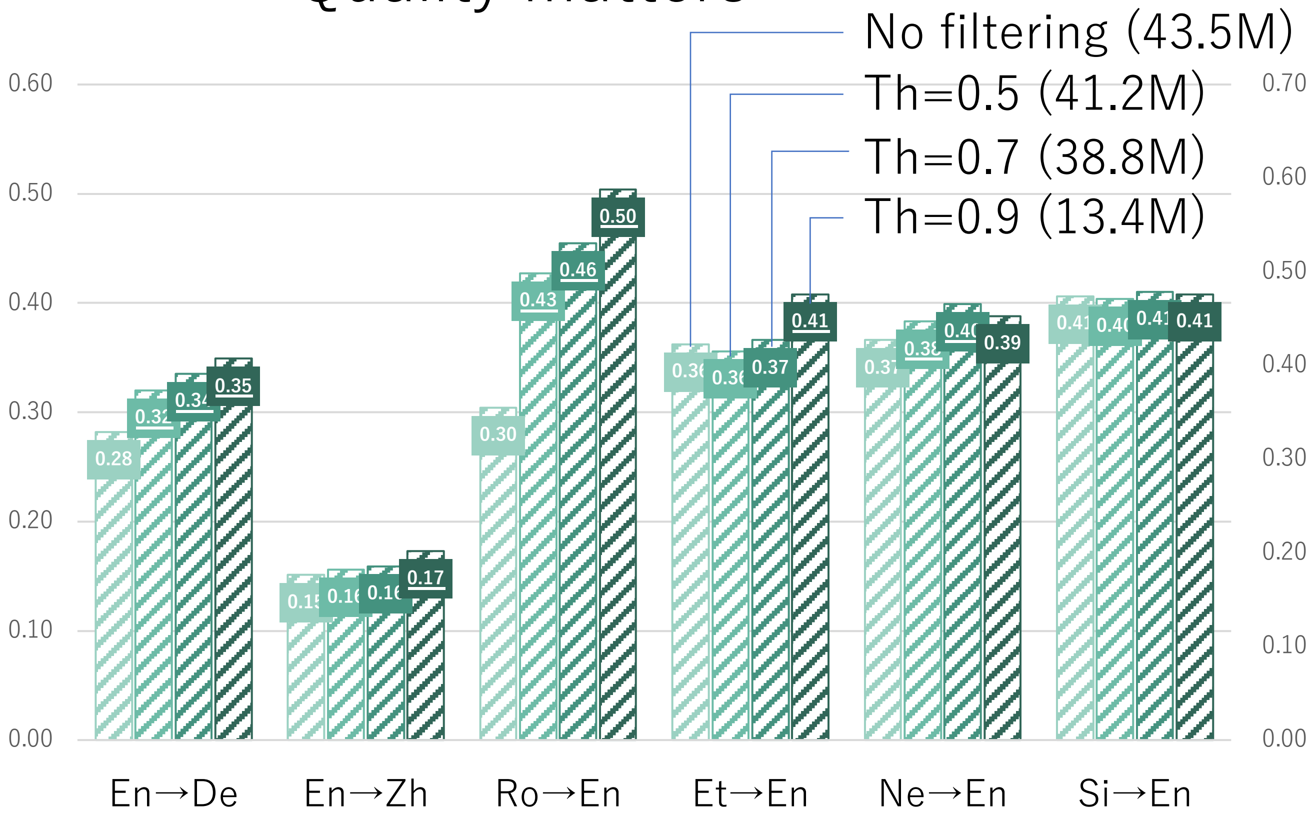
- Bilingual parallel corpora may contain noise
  - i.e., sentence pairs that are less likely to be translation
- We investigated the impact of the quality of parallel data as well as the quality of synthetic TQE data
  - Step 1. Computed a similarity score for each sentence pair
    - Cosine similarity between sentence embeddings based on LaBSE [Feng+ 2022]
  - Step 2. Filtered out pairs having a similarity lower than a pre-determined threshold
    - e.g., with a threshold of 0.5, only 60% of Ro→En pairs were retained
  - Step 3. Train models on filtered data



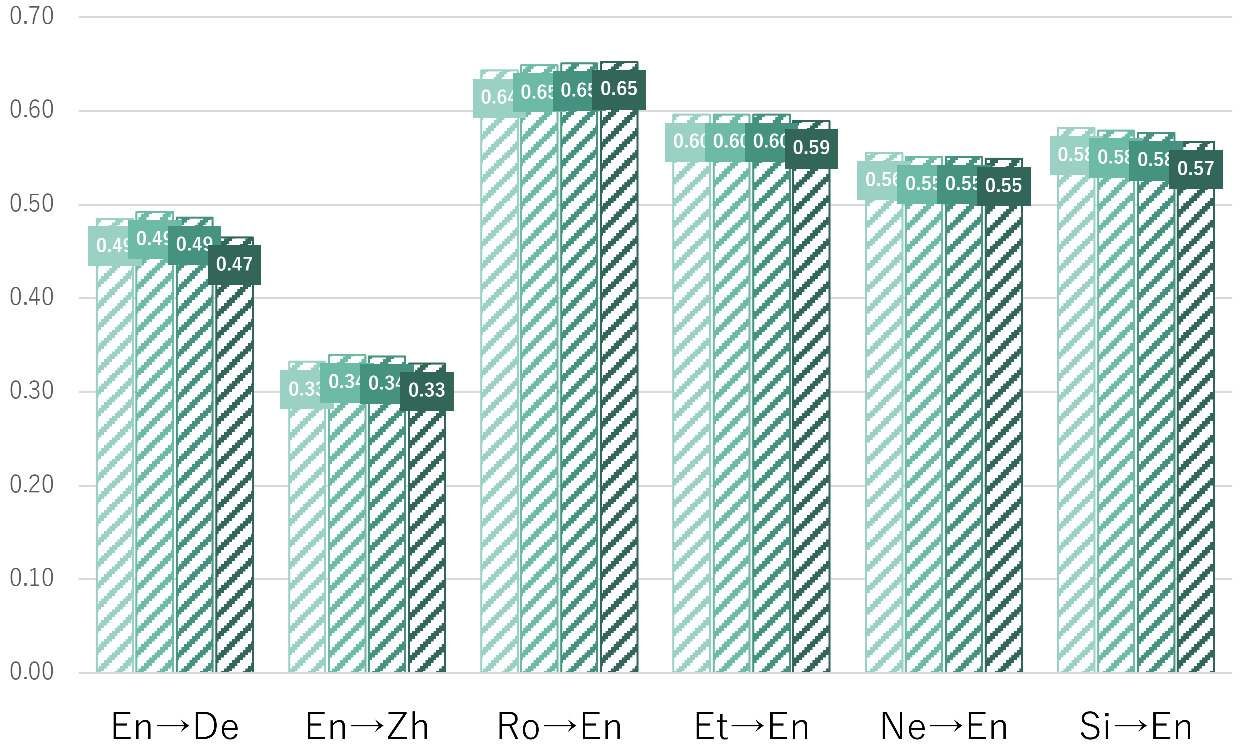
[Feng+ 2022] Feng et al. Language-agnostic BERT Sentence Embedding. In Proc. of ACL, 2022.

Impact of Synthetic Data Quality (Results)

- Pseudo-supervised models
  - Aggressive filtering of the parallel corpus led to higher MCCs
    → Quality matters



- Fine-tuned models
  - Filtering brought only a small impact



Conclusion

- We proposed to apply OT to determine pseudo-quality labels in synthetic data for word-level TQE
- Experimental results
  - OT-based labels better guide pre-training on a synthetic TQE data and lead to higher MCC in word-level TQE
  - Our method achieved consistently better results for pseudo-supervised settings as well as zero-shot translation directions

- Future work
  - Finer-grained hyper-parameter optimization (e.g., λ_m for each segment)
  - Labeling source words