# Cross-Lingual Summarization with Pseudo-Label Regularization

**Thang Le**
VinAI Research
v.thangld16@vinai.io

## Abstract

Cross-Lingual Summarization (XLS) aims to summarize a document in the source language into a condensed version in the target language, effectively removing language barriers for non-native readers. Previous approaches, however, have the same limitation that only a single reference (gold summary) is exploited during model training, making the base model exposed to an underrepresented hypothesis space since the actual number of possible hypotheses is exponentially large. To alleviate this problem, we present a study adopting pseudo-labels in regularizing standard cross-lingual summarization training. We investigate several components leading to the gains in regularization training with verified experiments involving 8 diverse languages from different families. Conclusively, we show that pseudo-labeling is a simple and effective approach that significantly improves over standard gold reference training in XLS.

## 1 Introduction

With the rise of massive information (Acharjya and KauserAhmed, 2022), summarization plays an essential part in absorbing and processing the emerging events in a timely and efficient manner (Syed et al., 2022). Given the existing large number of natural languages (Qin et al., 2024), the needs for providing information access to widespread groups of audience across nations have necessitated the construction of tasks beyond the monolingual setting (Shen et al., 2018). To facilitate this development, CROSS-LINGUAL SUMMARIZATION (XLS) was introduced where an article in a source language can be summarized to a condensed version in a different target language, effectively conveying key information to non-native readers (Duan et al., 2019).

Among modeling solutions for XLS, a straightforward approach was to adopt pipeline-based systems comprising of translation and summarization networks (Zhu et al., 2019). It is, however, not particularly appealing for several reasons including error propagation (Takase and Okazaki, 2022), loss of context and structure (Wang et al., 2023a), and the incurred extra computations along with reduced inference speed. Recent research have moved on to adopt end-to-end approaches where a single network executes the whole task (Zhu et al., 2019). To empower the cross-lingual performance, existing methods either scale up the training process (Xu et al., 2020; Wang et al., 2023b) or bootstrap learned representations through auxiliary objectives and features (Bai et al., 2021; Li et al., 2023; Jiang et al., 2022). Despite promising improvements, these methods have the same limitation in that for each training sample, the summarization network only gets exposed to a single gold reference. This mistakenly causes the network to learn skewed, deterministic one-point (Liu et al., 2022) and highly uncalibrated summary distributions (Zhao et al., 2023) that easily suffer from error accumulations such as exposure bias (Arora et al., 2022; Xie et al., 2023). Seeking prospective remedies, in this work, we study the training of end-to-end XLS networks in the presence of pseudo-labels i.e. pseudo-references with probability labels obtained from teacher models. In particular, we analyze several important components contributing to the performance of models regularized with pseudo-labels and present these findings as reference practices for future works. To inspect the effectiveness of this training recipe, we conduct empirical experiments spanning 8 typologically diverse languages coupled with comparisons against existing baselines.

In summary, our contributions can be listed as follows:

- We conduct a study on neural cross-lingual summarization with pseudo-label regularization. Particularly, we inspect factors contributing to the regularization's effectiveness to-

*Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4644–4677
June 16-21, 2024 ©2024 Association for Computational Linguistics

wards the model's performance, including the choice of teacher models and the construction of pseudo-references.

- To reinforce the study's findings, we conduct empirical experiments involving 8 distinct languages from different families with accompanying baseline comparisons where we observe significant improvements over standard gold reference training.

## 2 Related Works

**Cross-Lingual Summarization.** With the potential of reducing language barrier, cross-lingual summarization emerged as an active area of research in recent years (Shen et al., 2018; Zhu et al., 2019; Xu et al., 2020; Bai et al., 2021; Takase and Okazaki, 2022; Wang et al., 2022b; Fatima and Strube, 2023). Early works, due to the lack of parallel supervised corpora, often resort to pipeline methods comprising of separate translation and summarization networks (Shen et al., 2018). Zhu et al., 2019 first introduced a large-scale parallel corpus to experiment with end-to-end training where they also proposed a multi-task framework with improved performance. Subsequent works experimented with adversarial training (Cao et al., 2020), contrastive learning (Li et al., 2023), variational inference (Liang et al., 2022), knowledge distillation (Nguyen and Luu, 2022) and pre-training (Xu et al., 2020; Wang et al., 2022a, 2023b). These works, however, bear the same limitation in that they do not explore alternative hypotheses apart from the gold summary in each training sample which potentially leads to models' underperformance (Ranzato et al., 2016; Bengio et al., 2015) since they lacked exposure to the summary distribution which includes an exponentially large number of hypotheses. In contrast, we target the setting of training neural cross-lingual summarizers with additional supervision from pseudo-labels that are formed from pseudo-references with probability labels from teacher models.

**Pseudo-Labeling.** There have been ongoing research on adopting pseudo-labels in sequence-to-sequence training (Calderon et al., 2023). Liu et al., 2021 enriched the one-hot annotations with smoothed self-guidance under noised perturbations for summarization on English data. Shleifer and Rush, 2020 further examined replacing the ground truth target with the teacher's summary. Wang et al., 2021 additionally integrated selective mechanisms in forming training data points for translation tasks. Duan et al., 2019 relied on pseudo-sources and pseudo-labels to overcome scarcity in parallel samples for cross-lingual sentence summarization. Li et al., 2023 utilised soft-labels for enforcing cross-lingual consistency learning. Nevertheless, these works either do not consider multiple references during training (Duan et al., 2019; Li et al., 2023) or strictly focus on monolingual generation (Calderon et al., 2023) and machine translation (Zheng et al., 2018; Khayrallah et al., 2020). This leaves a research gap in cross-lingual summarization where the use of multiple training references remains neglected - which our work aims to fill in.

## 3 Background

### 3.1 Neural Cross-Lingual Summarization

Given a source document $D_A$ in the source language $A$, a neural cross-lingual summarizer needs to produce the summary $S_B$ in the target language $B$. Denote the target sequence's tokens as $[y_0, y_1..y_{|S_B|}]$, the training objective (negative log-likelihood) is defined as: $\mathcal{L}_{NLL} = -\sum_{t=1}^{|S_B|} log(P_\theta(y_t|y_{<t}, D_A))$, where $|S_B|$ denotes the length of the output summary and $\theta$ represents the model's parameters.

Since existing corpora only provide a single gold summary per sample, the ground truth label for $P_\theta(y_t|y_{<t}, D_A)$ is typically a $|V|$-dimensional one-hot vector, where $|V|$ is the vocabulary size. This leads to the underlying model being trained with an under-explored hypothesis space, as the number of possible hypotheses are exponentially large when in fact only one of them manifests during the training process. To alleviate this problem, we can integrate additional hypotheses (pseudo-references) as regularization during the model's training.

### 3.2 Pseudo-Labeling

To construct pseudo-references, a straightforward approach is to adopt decoding outputs from trained summarizers. This also simulates the model's behavior at inference time which helps ease the negative impact of exposure bias (Bengio et al., 2015; Arora et al., 2022) - a gap that arises because models are conditioned on ground truth tokens during training but have to rely on self-generated tokens during inference.

Assuming access to a teacher model $T_\phi$ which maps each pair $(D_A, Y')$ to a sequence of probability vectors $\{P_\phi(Y'_t|Y'_{<t}, D_A)\}$

(with $Y'$ being the pseudo-reference), we can define the regularization loss as the KL-divergence between the base model's prediction and the pseudo-label: $\mathcal{L}_{reg}(D_A, Y') = KL(P_\theta(Y'_t|Y'_{<t}, D_A)||P_\phi(Y'_t|Y'_{<t}, D_A))$. To incorporate multiple pseudo-labels for a single sample, we simply take the average over each individual loss: $\mathcal{L}_{reg} = \frac{1}{K}\sum_{i=1}^{K} \mathcal{L}_{reg}(D_A, Y^{i'})$.

The final training loss thus becomes: $\mathcal{L} = \mathcal{L}_{NLL} + \xi\mathcal{L}_{reg}$, where $\xi$ is the hyperparameter controlling the regularization effect[1].

## 4 Empirical Analysis

In this section, we investigate the choice of the teacher models for obtaining the probability labels as well as the forming of pseudo-references, both of which make up the pseudo-labels for training. Additionally, we examine their effects on the XLS model's performance and further see how they fare against the gold labels. Particularly, we focus on the following research questions:

- **Q1: Which model, or combination of models, should we adopt to label the pseudo-references ?** (4.1, 4.2, 4.4)
- **Q2: Should we use stochastic or static probability labels ?** (4.3)
- **Q3: Which groups of pseudo-references work better ?** (4.5)
- **Q4: Do the number of pseudo-references matter ?** (4.6)
- **Q5: Can soft probability labels replace one-hot annotations in gold summaries ?** (4.7)
- **Q6: How competitive are pseudo-labels ?** (4.8)

**Task Setup.** For experimental purposes, we use the WIKILINGUA[2] dataset (Ladhak et al., 2020 ) with a focus on 8 languages: English(EN), Vietnamese(VI), Japanese(JA), Chinese(ZH), Arabic(AR), Korean(KO), Russian(RU) and Turkish(TR). These languages come from different families (Table 1) and possess distinct morphological and/or topological characteristics which ensure the experiment's coverage. For base architecture, we mainly employ the MBART-50[3] model (Tang et al., 2021) unless explicitly mentioned otherwise. Additionally, we use beam search with a decoding beam of 32 to construct the pseudo-reference pool from monolingual models (we explain this decision later on). As simultaneously forwarding $K$ pseudo-references through the model would require duplicating the encoder hidden states by $K$ times, which would increase the memory usage whereas sequentially forwarding these pseudo-references (for each sample) reduces model training speed. In our experiments, we uniformly sample $K$ hypotheses from the pseudo-reference pool at each iteration and simultaneously forward these hypotheses[4]. Thus, the base model can get exposed to different sets of pseudo-references at each epoch while keeping memory consumption at a sustainable level.

For each training run, we use the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of $1e-5$ and a linear decay scheduler for a maximum of 300 000 steps. We use a batch size of 4 for standard training and 2 when adopting pseudo-labels. For evaluation, we primarily use the ROUGE metrics (Lin, 2004). In the multilingual setting, previous works either utilize the MULTILINGUAL ROUGE toolkit[5] for language-specific processing (Hasan et al., 2021; Wang et al., 2022a; Bhattacharjee et al., 2023) or the SENTENCEPIECE model for language-agnostic tokenization (Vu et al., 2022; Li and Murray, 2023; Clark et al., 2023), which we hereinafter refer to as M-ROUGE and SP-ROUGE, respectively. Upon measuring these metrics with human judgements from the MULTISUMMEVAL dataset (Koto et al., 2021), we find that SP-ROUGE generally attains better correlation[6]. We therefore report results obtained with the SP-ROUGE metric[7]. For presentation's sake, we focus on the ROUGE-L variant and include the remaining results in the Appendix. Experiments were implemented with the PyTorch (Paszke et al., 2019) and Transformer (Wolf et al., 2019) frameworks, and executed on an A100 GPU.

[1] We set $\xi = 3.2$ in all experiments. We obtained this value through tuning $\xi$ in the range $[0.0, 4.0]$ based on validation performance of an initial language pair (EN2JA)

[2] The dataset is publicly available under a Creative Common license: https://github.com/esdurmus/Wikilingua

[3] https://huggingface.co/facebook/mbart-large-50

[4] We choose $K = 2$ due to memory limit.

[5] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

[6] See Appendix C

[7] We use the SENTENCEPIECE model from MBART-50

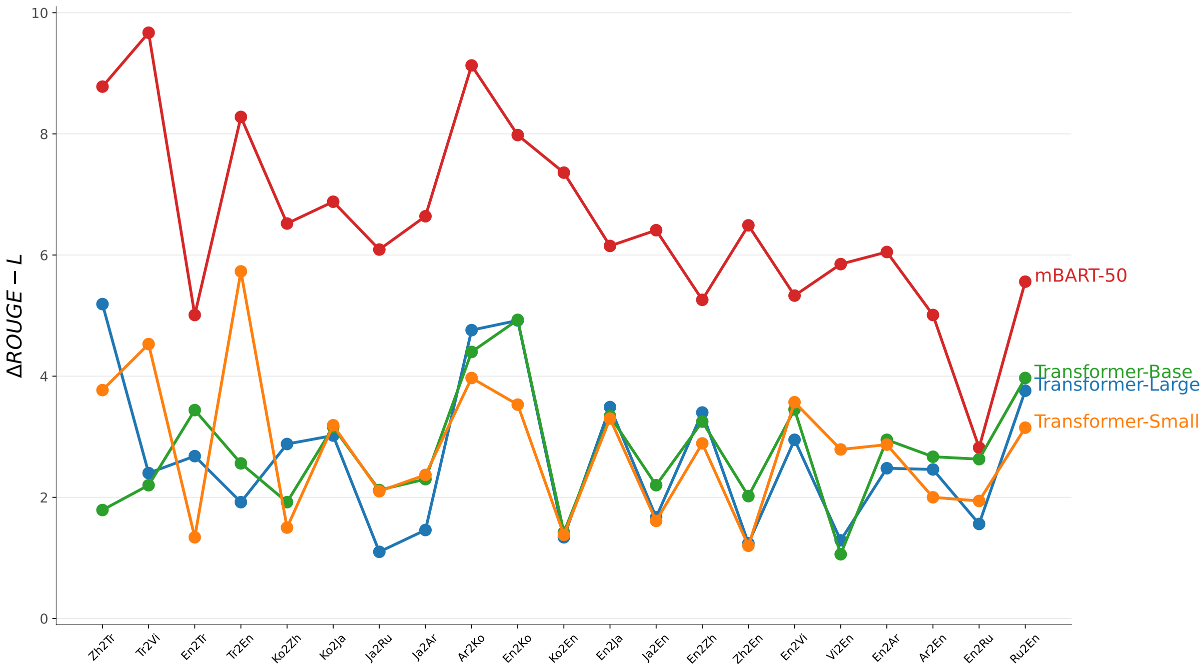

Figure 1: Characterizing the CROSS-LINGUAL / MONOLINGUAL Summarization Gap (VALIDATION)

| Code | Language | Family |
|---|---|---|
| En | English | Indo-European: West Germanic |
| Vi | Vietnamese | Austroasiatic: Viet-Muong |
| Ja | Japanese | Japonic |
| Zh | Chinese | Sino-Tibetan |
| Ar | Arabic | Afro-Asiatic: Central Semitic |
| Ko | Korean | Koreanic |
| Ru | Russian | Indo-European: East Slavic |
| Tr | Turkish | Turkic |

Table 1: List of languages included in the study

### 4.1 Monolingual Summarizers are Good Teachers

Given parallel corpora, each document $D_A$ in the source language is also accompanied with a document $D_B$ in the target language. This means that we can use a monolingual summarization model $M_B$ to label the pseudo-reference $Y'_B$. As the monolingual model does not need to perform cross-language alignment or translation operations and can exploit easier shortcuts such as copying or rephrasing (Song et al., 2020), we argue that monolingual summarizers are approriate labelers that can assign high-quality probability labels.

To validate this argument, we compare the performance between the cross-lingual and monolingual summarizers over 21 cross-language directions where the samples used to train the cross-lingual and monolingual models are identical, and only the source and target languages vary. In other words, the monolingual model is trained on exactly the same data as the cross-lingual model, but with the source language being switched to the target language.

Specifically for this experiment, we examine both models trained from scratch and those initialized with a pre-trained language model. In particular, we train three TRANSFORMER (Vaswani et al., 2017) architectures from scratch, each containing $4/8/12$ layers in the encoder-decoder stacks (hereinafter termed SMALL/BASE/LARGE). Parameters were initialized with the Xavier Uniform initialization (Glorot and Bengio, 2010). We use a hidden state size of $1024$ and a feedforward dimension of $4096$. Dropout rates are set to $0.1$ and layer normalization is placed inside the residual blocks (Pre-LN) (Xiong et al., 2020). For pre-trained architecture, we fine-tune the MBART-50 model.

We present the results in Figure 1 (VALIDATION). For each direction, we plot the value $\Delta P = P_{Mono} - P_{Cross}$ as the performance gap between the two summarizers, with $P$ being ROUGE-L. We can see that the monolingual summarizer always performs better despite being trained on the same samples. Noticably, the gap is quite similar across architecture scales (SMALL/BASE/LARGE) but significantly rises for the MBART-50 model. This is most likely due to the pre-training stage which involves monolingual denoising objectives making MBART-50 much better at monolingual tasks while being less likely to conduct cross-lingual generation.

**TL;DR:** *At equivalent training scales, monolingual models are shown to exhibit stronger summarization performance than cross-lingual models.*

### 4.2 Ensemble Teacher

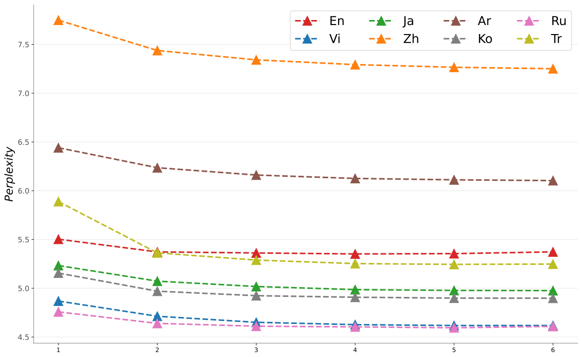

Figure 2: Ensemble Set Perplexity (VALIDATION)

We next investigate whether we can improve the quality of pseudo-labels by combining probability predictions from several teacher models. For this purpose, we train six distinct monolingual models for each language and combine them to form ensemble sets of varying sizes. For each ensemble size, we utilize the combination with the lowest perplexity score. In particular, we average each model prediction in the ensemble set to get the final probabilities.

We show the results in Figure 2 (VALIDATION). Each circle denotes the perplexity score of the best ensemble set of the given size for each language. Results show we can readily obtain higher-quality probability labels via taking the mean predictions of different models, as indicated by lower perplexity scores. However, using a large number of models will incur non-trivial memory overhead. For example, storing a MBART-50 model in single precision (FP32) would require 3.5GB of VRAM. To keep memory consumption at a tolerable level, we only use an ensemble size of 3 in the succeeding experiments.

**TL;DR:** *Averaging teacher predictions yields higher-quality probability labels.*

### 4.3 Teacher Dropout

By applying dropout on the teacher models, we can obtain stochastic probability labels that change at each iteration. We investigate the difference

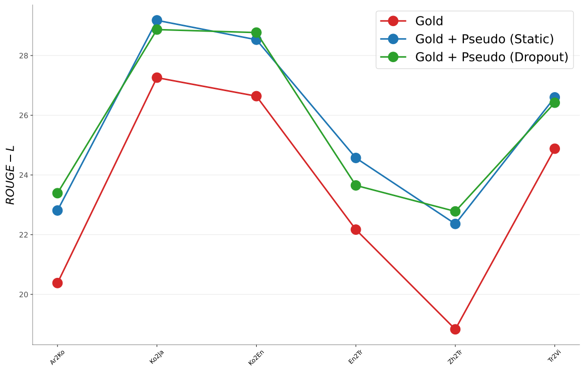

Figure 3: Dropout on probability labels (VALIDATION)

between training with these stochastic probability labels (without dropout) and static ones in Figure 3 (VALIDATION). Although the probability labels with dropout are supposed to be noisier, we find the resulting models to perform better than those with static probability labels on 3/6 directions, but this does not depict significant difference on average[8]. Still, we apply dropout in the remaining experiments hoping to avoid excessive overfitting.

**TL;DR:** *Apparently, there is no significant difference on average between stochastic and static probability labels.*

### 4.4 Teacher Comparison

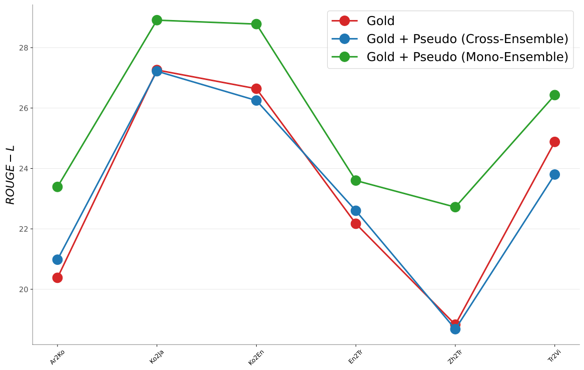

Figure 4: Teacher Ensemble Performance (VALIDATION)

Although we showed that monolingual summarizers produced higher-quality pseudo-labels, it is important to verify that these pseudo-labels also give rise to better performance. We thus conduct regularization with either an ensemble of monolingual or cross-lingual teachers. The results are shown in Figure 4 (VALIDATION). Here the monolingual teachers clearly play an important part in

[8] See Appendix D

empowering the base summarizer, achieving better scores than the other two model variants. Meanwhile, the cross-lingual teachers often do not yield significant improvement and even lead to negative results in some directions.

**TL;DR:** *Pseudo-labels from monolingual models consistently improve cross-lingual summarization performance.*

## 4.5 Pseudo-Reference Quality

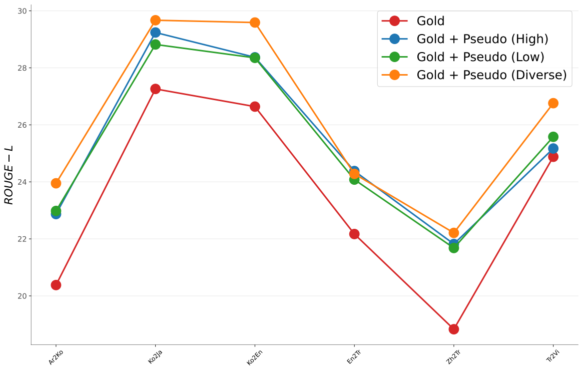

Figure 5: Pseudo-references properties (VALIDATION)

We next study whether the properties of the pseudo-references would affect regularization training. From the pseudo-reference pool (constructed with beam search), we each sample 16 hypotheses with highest and lowest ROUGE scores[9], respectively. Additionally, we use diverse beam search (Vijayakumar et al., 2016) to obtain 16 hypotheses with higher diversity. These hypotheses are then used in regularization training.

Results are shown in Figure 5 (VALIDATION). We can see that the quality matters since training with the uppermost 16 pseudo-references is better than with the lowermost on $4/6$ directions. Besides, variety also contributes to the regularization's outcomes as diverse pseudo-references facilitate better results than the remaining two on $5/6$ directions. Nevertheless, the runtime of diverse beam is much longer than that of standard beam. For example, generations of 16 diverse candidates for each training example in the AR2KO direction are 2.3 times slower than alternatively using standard beam while taking up significantly more GPU memory. Thus, it would make experiments less scalable and reduce the amount of coverage we could afford. As a result, we only adopt standard beam decoding in other pseudo-labeling experiments.

**TL;DR:** *High-quality pseudo-references help, but diversity is also beneficial.*

## 4.6 Size of Pseudo-Reference Pool

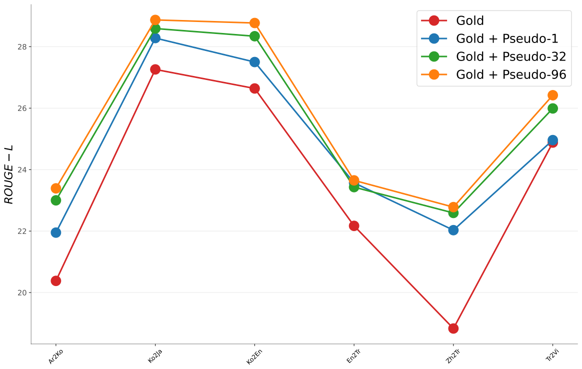

Figure 6: Number of pseudo-references (VALIDATION)

We next examine whether the size of the pseudo-reference pool affects the regularization's effectiveness. We randomly subsample subsets of size $1/32/96$ from the original pseudo-reference pool and use these subsets as new pools in regularization training. Results are shown in Figure 6 (VALIDATION). We observe that the larger the pool's size is, the better the results are. This means the number of pseudo-references can have a significant impact on the regularized model's performance. Noticeably, simply training with one additional pseudo-reference can already bring forth gains in $6/6$ directions, further highlighting the necessity of regularization training.

**TL;DR:** *A larger size of the pseudo-reference pool works better.*

## 4.7 Pseudo-Labeling on Gold Summary

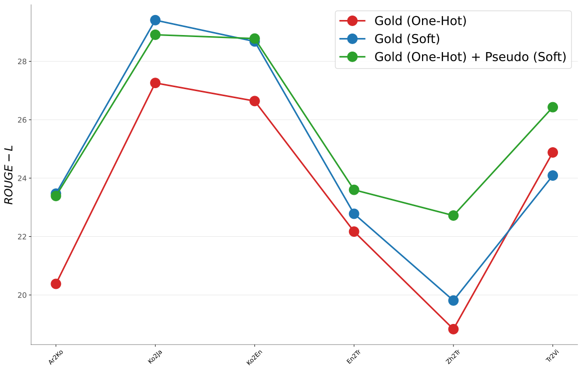

Figure 7: Soft Labeling on Gold Summary (VALIDATION)

Given the effectiveness of pseudo-labeling, we are also interested in whether using teacher models

[9] Here we use the average of ROUGE-1/2/L as the selection criterion to balance across ROUGE variants

to label the gold summaries would turn out better than using the default one-hot labels. Results are shown in Figure 7 (VALIDATION). Surprisingly, soft labeling also works well when applied on the gold summaries, yielding improvements on 5/6 directions over the one-hot annotations. This further corroborates the pseudo-probability labels' qualities. Still, combining the one-hot gold summaries with soft pseudo-labels (multiple pseudo-references) generally works better, attaining highest scores on 4/6 directions.

**TL;DR:** *Annotating the gold summaries with pseudo-probability labels vastly improves performance, but combining the one-hot gold summaries with pseudo-labels works better.*

### 4.8 Regularization as the sole objective

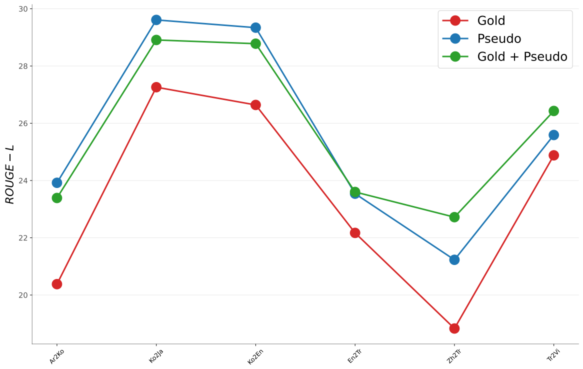

Figure 8: Training with only pseudo-labels (VALIDATION)

We further investigate training with only pseudo-labels (no gold summaries). Results are shown in Figure 8 (VALIDATION). Training with only pseudo-labels turn out to be better than standard training with gold labels on 6/6 directions, and only worse than training with both types of labels on 3/6 directions. Nevertheless, we observe that on average, better results are obtained with both types of labels combined[10].

**TL;DR:** *Only using pseudo-labels outperforms gold label training, but combining both types of labels works better.*

## 5 Benchmarking

In this section, we study the effectiveness of pseudo-label regularized models on three settings: FULLY SUPERVISED LEARNING, FEW-SHOT LEARNING and PARAMETER-EFFICIENT FINE-TUNING. We follow lessons derived from the previous section for the teacher models' choice as well as the construction of pseudo-references. In particular, we use an ensemble of 3 monolingual models for teacher labeling. For pseudo-reference, we use beam decoding with a beam size of 32 from three monolingual models for each language to construct the pseudo-reference pool with a total size of 96. Unless explicitly mentioned otherwise, we use the MBART-50 language model as the base architecture, similar to the previous section. For further implementation details, we refer readers to Appendix B.

### 5.1 Fully Supervised Learning

**Baselines.** We compare the regularized models with the following baselines:

- **Trans-Sum** - A pipeline where the input article is first translated to the target language with a translation network, then summarized with a monolingual summarization network to obtain the cross-lingual summary
- **Sum-Trans** - A pipeline where the input article is first summarized in the source language with a monolingual summarization network, then translated with a translation network to obtain the cross-lingual summary
- **Gold** - The standard end-to-end model trained using the gold summaries with one-hot annotations (Zhu et al., 2019)
- **Gold + OTSum** - An end-to-end model that is further equipped with a knowledge distillation loss based on the optimal transport distance (Nguyen and Luu, 2022)
- **Gold + Many2Many** - An end-to-end model that is trained using parallel samples from all cross-lingual directions (Wang et al., 2023b; Bhattacharjee et al., 2023). This model is included primarily for reference purposes, as it requires vastly more training data than other models.

**Automatic Evaluation.** We depict results on the TEST split with automatic evaluation using ROUGE-L in Figure 9. Average scores across 21 directions and statistical significance are reported in Table 2. Compared to the baselines, adopting pseudo-labels clearly improves performance on many directions. Particularly, compared to standard training (GOLD),

[10] See Appendix D

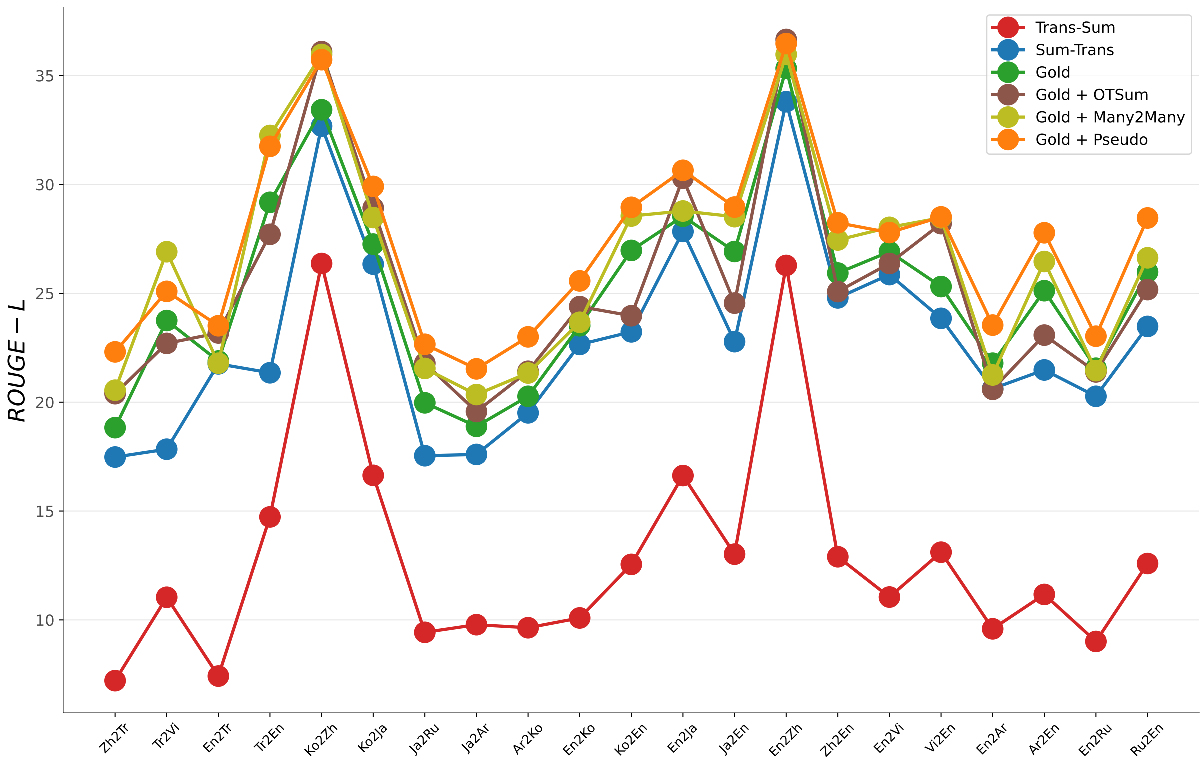

Figure 9: Fully Supervised Training (Test)

Gold+Pseudo yields significant improvements in $21/21$ directions. Even when compared to Gold+Many2Many which uses 16X-274X more parallel data, Gold+Pseudo still performs significantly better in $14/21$ directions. Overall, we see a substantial rise of 2.2 points in ROUGE-L compared to the base Gold model, validating the regularization's effectiveness.

| | *Average* | $p < 0.05$ |
|---|---|---|
| Trans-Sum | 12.87 | 21/21 |
| Sum-Trans | 22.99 | 21/21 |
| Gold | 25.11 | 21/21 |
| Gold + OTSum | 25.31 | 16/21 |
| Gold + Many2Many | 26.40 | 14/21 |
| **Gold + Pseudo** | **27.31** | - |

Table 2: Average ROUGE-L scores across 21 directions and the number of directions where Gold+Pseudo achieves significantly higher scores ($p < 0.05$) than the according baseline.

**Manual Evaluation.** Since automatic metrics have certain intrinsic limitations (Koto et al., 2021; Clark et al., 2023), we further conduct human evaluation to assess the quality difference between trained models. Particularly, we select two cross-language directions: Russian-to-English (Ru2En) and English-to-Vietnamese (En2Vi). From the test set of each direction, we randomly selected 50 document-summary pairs in the target language along with three model-generated cross-lingual summaries[11]: Gold, Gold+OTSum, Gold+Pseudo. As participants, we invited three professional English speakers and three native Vietnamese speakers to rank the generated summaries in terms of two core aspects: *informativeness* (to which extent a summary captures the main content) and *fluency* (how well-formed a summary is in the target language). Each sample was examined by three participants and each model received a score of 1-3 for each rating according to its rank. Av-

| | *Ru2En* | | *En2Vi* | |
|---|---|---|---|---|
| | `IF` | `FL` | `IF` | `FL` |
| Gold | 1.88 | 2.02 | 1.87 | 2.23 |
| Gold + OTSum | 1.93 | 2.05 | 1.91 | 2.33 |
| **Gold + Pseudo** | **2.59** | **2.60** | **2.41** | **2.55** |

Table 3: Human evaluation on the Ru2En and En2Vi directions. `IF` and `FL` each denotes *informativeness* and *fluency*, respectively.

[11] Here we exclude the Gold+Many2Many model as the amount of parallel data used to train it vastly differs from the others, making it less comparable

eraged results are shown in Table 3. We observe that the ratings are significantly higher ($p < 0.05$) for GOLD+PSEUDO in the two directions on both categories compared to the remaining two models, indicating that the summaries produced by GOLD+PSEUDO were also preferred more by human participants.

**Other Architectures.** Beyond MBART-50, we additionally experiment with the PISCES model (Wang et al., 2023b) and a TRANSFORMER model[12] trained from scratch. We train separate monolingual models with these architectures while re-using the pseudo-reference pools from the previous experiment to form the pseudo-labels, which we then apply to regularize cross-lingual summarizers trained with these architectures. We examine the results in two directions: AR2KO and KO2JA (Figure 10). We observe that pseudo-labels are effective on these two model types as well, consistently improving upon the GOLD models. This means that pseudo-labels' utility is not limited to any single model architecture.

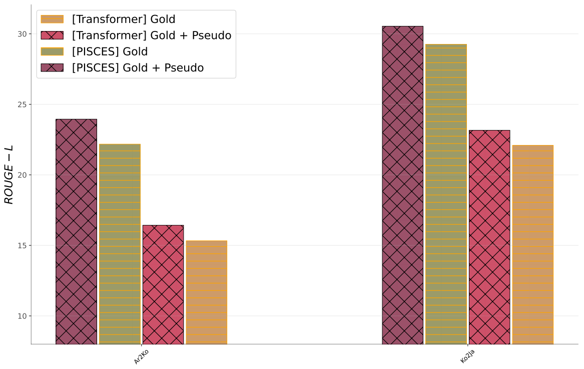

Figure 10: FULLY SUPERVISED TRAINING with PISCES and TRANSFORMER as the base models (TEST)

## 5.2 Few-shot Learning

We next investigate pseudo-label's potency under few-shot settings, where we only use as little as 50/100/300 parallel training samples in each direction (Figure 11). Even under these extreme scenarios, we observe that pseudo-labels still effectively reinforce the base models' performance.

## 5.3 Parameter-Efficient Fine-Tuning

Although PARAMETER-EFFICIENT FINE-TUNING boosts training efficiency, it often limits model's capabilities (Song et al., 2024). We thus examine whether pseudo-labels can help amend this limitation. In particular, we apply LOW-RANK ADAPATION (LORA) (Hu et al., 2022) to the base MBART-50 model and trains it with pseudo-labeling similar to previous sections. For LORA hyperparameters, we use a rank $r$ of 8, an $\alpha$ value of 32 and a dropout rate of 0.1 with modules only inserted after the query and value projection matrices. Results are shown in Figure 12. In this setting, we also find pseudo-labeling effective, consistently improving over the base [LORA] GOLD model.

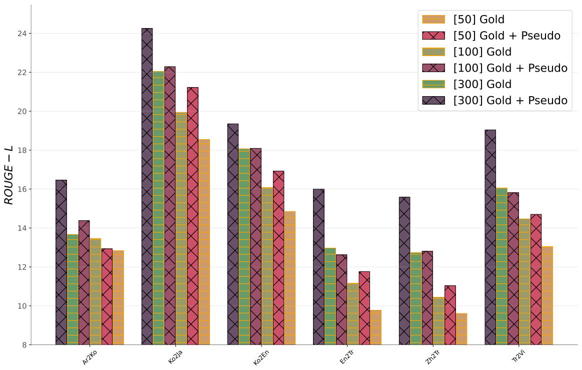

Figure 11: FEW-SHOT LEARNING with 50/100/300 training samples (TEST)

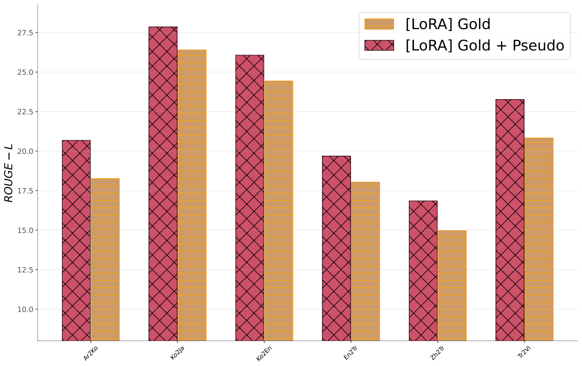

Figure 12: LOW-RANK ADAPTATION results (TEST)

# 6 Conclusion

In this paper, we study the effectiveness of pseudo-label regularization in standard neural cross-lingual summarization training. We conduct empirical experiments involving 8 diverse languages from different families and study the different components affecting the end performance of regularized models. We further validate the regularization's usage in three distinct learning settings. The results show that pseudo-labeling is a simple but effective companion complementary to the standard training procedure. We hope our study will prove useful for future practitioners working on this task.

[12] Here we use the LARGE scale as in Section 4.1

## Limitations

In this paper, we have only shown that pseudo-labeling is effective in supervised settings (either *high-* or *low-resource* scenarios). For many cross-language directions, parallel corpora might not exist (Liu et al., 2020) and thus pseudo-labels cannot be readily applied in these situations. Nevertheless, should parallel translation data (or models) exist, it is possible to construct pseudo-parallel data samples (Shen et al., 2018) from which the pseudo-labels discussed in this work can be applied directly, which will be a prospective direction for future works.

In addition, the experiments in this work were conducted solely on the WIKILINGUA (Ladhak et al., 2020) dataset. As a result, we are not entirely certain whether the findings in this work would still generalize to other domains or datasets. For example, would the monolingual models still provide high-quality pseudo-labels ? Or would the pseudo-references' properties examined in Section 4 still hold the same level of importance ? These questions would require further verifications to answer.

Besides, summarization evaluation involves several distinct aspects, some of which we do not yet fully cover in this work. For example, while we made efforts to quantify *informativeness* and *fluency*, we did not explicitly quantify *attribution* (Clark et al., 2023) or *hallucination* (Aharoni et al., 2023), which future works should also pay attention to.

## Acknowledgment

We thank the anonymous reviewers and meta reviewer for their valuable feedback and suggestions.

## References

Debi Prasanna Acharjya and P KauserAhmed. 2022. A survey on big data analytics: Challenges, open research issues and tools. *International Journal for Research in Applied Science and Engineering Technology*.

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 700–710, Dublin, Ireland. Association for Computational Linguistics.

Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6910–6924. Association for Computational Linguistics.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1171–1179.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. Crosssum: Beyond english-centric cross-lingual summarization for 1, 500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2541–2564. Association for Computational Linguistics.

Nitay Calderon, Subhabrata Mukherjee, Roi Reichart, and Amir Kantor. 2023. A systematic study of knowledge distillation for natural language generation with pseudo-target training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14632–14659. Association for Computational Linguistics.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6220–6231. Association for Computational Linguistics.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3162–3172. Association for Computational Linguistics.

Mehwish Fatima and Michael Strube. 2023. Cross-lingual science journalism: Select, simplify and rewrite summaries for non-expert readers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1843–1861. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 249–256. JMLR.org.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Shuyu Jiang, Dengbiao Tu, Xingshu Chen, Rui Tang, Wenxian Wang, and Haizhou Wang. 2022. Cluegraphsum: Let key clues guide the cross-lingual abstractive summarization. *CoRR*, abs/2203.02797.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–89, Online. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 801–812, Online. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen R. McKeown. 2020. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4034–4048. Association for Computational Linguistics.

Peiyao Li, Zhengkun Zhang, Jun Wang, Liang Li, Adam Jatowt, and Zhenglu Yang. 2023. ACROSS: an alignment-based framework for low-resource many-to-one cross-lingual summarization. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2458–2472. Association for Computational Linguistics.

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12461–12476, Toronto, Canada. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022. A variational hierarchical model for neural cross-lingual summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2088–2099. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Sheng Shen, and Mirella Lapata. 2021. Noisy self-knowledge distillation for text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–703, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yixin Liu, Pengfei Liu, Dragomir R. Radev, and Graham Neubig. 2022. BRIO: bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2890–2903. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11103–11111. AAAI Press.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Shiqi Shen, Yun Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(12):2319–2327.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation. *CoRR*, abs/2010.13002.

Haobo Song, Hao Zhao, Soumajit Majumder, and Tao Lin. 2024. Increasing model capacity for free: A simple strategy for parameter efficient fine-tuning. In *The Twelfth International Conference on Learning Representations*.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Ren Liu, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. 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 8902–8909. AAAI Press.

Ayesha Ayub Syed, Ford Lumban Gaol, Alfred Boediman, Tokuro Matsuo, and Widodo Budiharto. 2022. A survey of abstractive text summarization utilising pretrained language models. In *Intelligent Information and Database Systems - 14th Asian Conference, ACIIDS 2022, Ho Chi Minh City, Vietnam, November 28-30, 2022, Proceedings, Part I*, volume 13757 of *Lecture Notes in Computer Science*, pages 532–544. Springer.

Sho Takase and Naoaki Okazaki. 2022. Multi-task learning for cross-lingual abstractive summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 3008–3016. European Language Resources Association.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6456–6466, Online. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023a. Understanding translationese in cross-lingual summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3837–3849. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022a. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7716–7729. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022b. A survey on cross-lingual summarization. *Trans. Assoc. Comput. Linguistics*, 10:1304–1323.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Towards unifying multi-lingual and cross-lingual summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15127–15143. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Jiawen Xie, Qi Su, Shaoting Zhang, and Xiaofan Zhang. 2023. Alleviating exposure bias via multi-level contrastive learning and deviation simulation in abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9732–9747. Association for Computational Linguistics.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020. Mixed-lingual pretraining for cross-lingual summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 536–541, Suzhou, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J. Liu. 2023. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Renjie Zheng, Mingbo Ma, and Liang Huang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3188–3197, Brussels, Belgium. Association for Computational Linguistics.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3052–3062. Association for Computational Linguistics.

# A Method Illustration

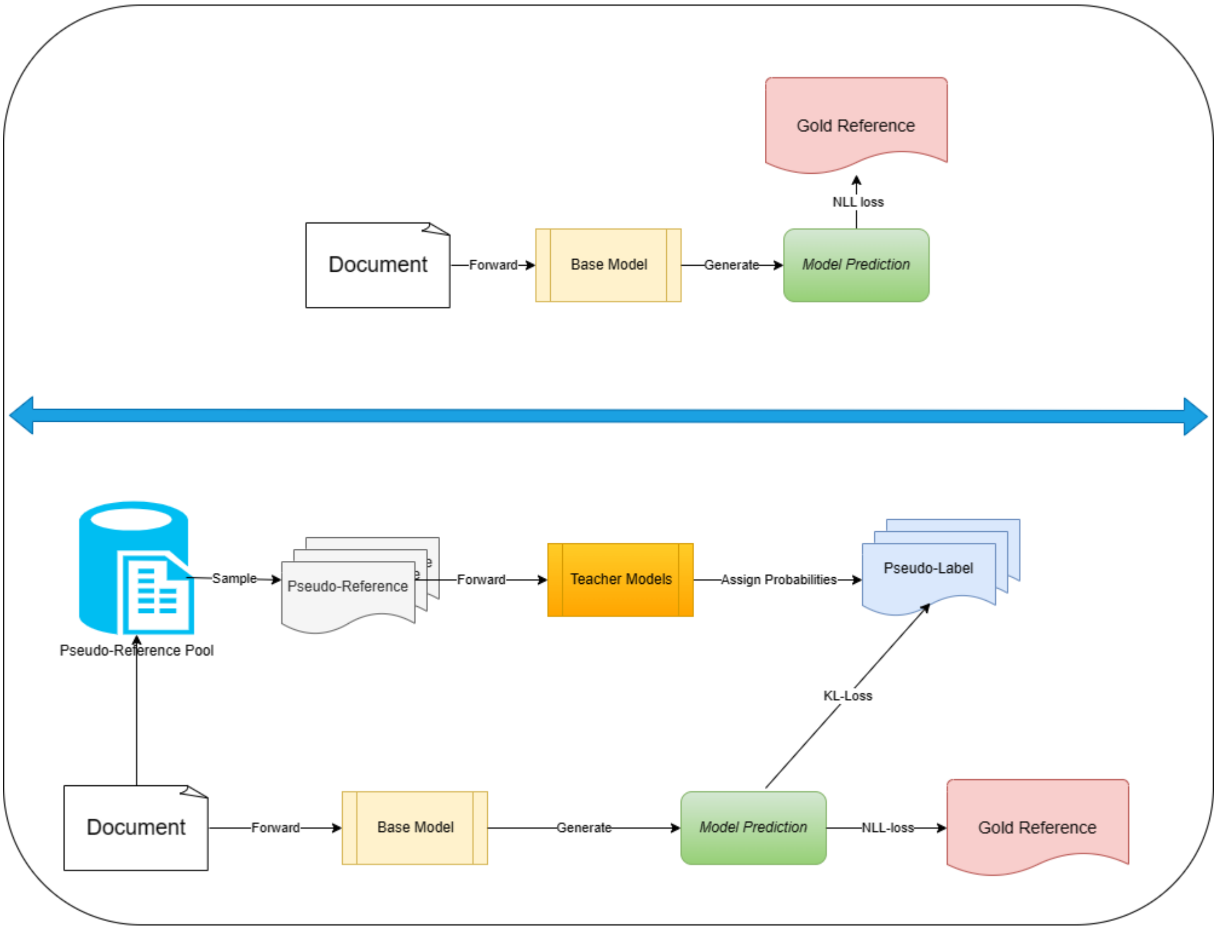

Figure 13: Illustrations of standard gold reference training (upper) and training with pseudo-labels (lower).

# B Benchmarking Details

For the two pipelines that require intermediate translation, we each fine-tuned one document-level and one summary-level MBART-50 translation networks for every cross-language direction using parallel samples from WIKILINGUA (Ladhak et al., 2020). For the summarization component, we re-used the trained monolingual summarizers. Regarding GOLD+OTSUM, Nguyen and Luu, 2022 originally initialized their framework with the encoder from MBERT (Devlin et al., 2019) and two TRANSFORMERS decoders, where the parameters corresponding to the teacher were first pre-trained and afterwards used to initialize the student, then jointly trained both the student and teacher models with the shared encoder and embedding layers during distillation. However, we find that this type of shared training severely impairs performance on MBART-50 (compared to GOLD) and that freezing the teacher model works better. Therefore, we kept the teacher model frozen, initialized the student separately and did not share parameters between the two networks during distillation in GOLD+OTSUM. For GOLD+MANY2MANY, we aggregated parallel samples from all cross-lingual directions possibly composed from the 8 languages[13] (Table 1) but did not include monolingual samples as in Wang et al., 2023b since we do not evaluate multilingual summarization. Training sampling was based on $p(D) \propto |D|^{\beta}$, where $p(D)$ is the probability of sampling from a given direction and $|D|$ is the number of training instances in that direction. We set $\beta = 1$ in our experiments. On training steps, we fine-tuned the GOLD+MANY2MANY model with a maximum of 600 000 steps due to large training data whereas other summarization models were fine-tuned with a maximum of 300 000 steps. During inference (for evaluation), we used a beam size of 8 with a length

[13] This yields $8 * 7 = 56$ directions

penalty of 1.0 and disabled any trigram repetition. The minimum and maximum generation lengths were set to 32 and 256, respectively. The maximum sequence lengths for the source article and the target summary were set to 1024 and 256, respectively.

## C Metric Correlation

| | *Focus* | | | | | *Coverage* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | En | Ru | Tr | Zh | Avg | En | Ru | Tr | Zh | Avg |
| SP-ROUGE-1 | 0.57 | 0.64 | 0.82 | 0.78 | 0.70 | 0.61 | 0.57 | 0.79 | 0.73 | 0.67 |
| M-ROUGE-1 | 0.59 | 0.61 | 0.83 | 0.76 | 0.69 | 0.62 | 0.52 | 0.78 | 0.68 | 0.65 |
| SP-ROUGE-2 | 0.53 | 0.65 | 0.79 | 0.70 | 0.67 | 0.54 | 0.56 | 0.75 | 0.65 | 0.62 |
| M-ROUGE-2 | 0.54 | 0.62 | 0.77 | 0.64 | 0.64 | 0.53 | 0.51 | 0.72 | 0.59 | 0.59 |
| SP-ROUGE-L | 0.55 | 0.62 | 0.82 | 0.78 | 0.69 | 0.57 | 0.52 | 0.77 | 0.73 | 0.65 |
| M-ROUGE-L | 0.55 | 0.59 | 0.82 | 0.17 | 0.53 | 0.58 | 0.49 | 0.77 | 0.12 | 0.49 |

Table 4: Pearson correlation scores of the two ROUGE variants (SP- and M-) with human judgements on four languages (English, Russian, Turkish, Chinese) from the Multi-SummEval dataset (Koto et al., 2021). We use ROUGE F1 scores for both *Focus* and *Coverage*. SP-ROUGE generally attains higher correlations than M-ROUGE.

# D Detailed Results

## D.1 ROUGE-1

| | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mBART-50 | 10.15 | 10.95 | 5.30 | 9.28 | 7.58 | 8.49 | 7.35 | 8.23 | 10.28 | 9.57 | 8.34 | 7.24 | 6.91 | 6.24 | 6.95 | 5.70 | 6.33 | 7.31 | 5.75 | 3.41 | 6.04 |
| Transformer-Large | 5.60 | 3.72 | 3.63 | 2.83 | 4.51 | 4.99 | 1.92 | 1.88 | 6.84 | 7.65 | 1.89 | 5.19 | 2.40 | 5.23 | 1.76 | 3.98 | 2.09 | 3.60 | 3.35 | 2.23 | 5.45 |
| Transformer-Base | 2.14 | 3.40 | 4.54 | 4.28 | 3.61 | 5.22 | 3.93 | 3.41 | 6.80 | 7.18 | 2.17 | 5.15 | 3.16 | 5.23 | 3.01 | 5.16 | 1.83 | 4.39 | 4.07 | 4.06 | 5.93 |
| Transformer-Small | 4.11 | 6.77 | 1.68 | 6.98 | 2.49 | 4.42 | 3.92 | 3.51 | 5.84 | 5.60 | 2.16 | 5.17 | 2.35 | 4.68 | 1.47 | 5.29 | 5.11 | 4.78 | 2.64 | 2.91 | 4.49 |

Table 5: Quantifying the cross-/mono-lingual summarization gap. Scores (validation set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.61 | 34.12 | 33.63 | 27.04 | 23.23 | 31.79 | 29.40 |
| Gold + Pseudo (Static) | 29.14 | 36.05 | 35.88 | 30.00 | 27.39 | 33.25 | 31.95 |
| Gold + Pseudo (Dropout) | 29.76 | 36.02 | 36.23 | 28.86 | 27.90 | 33.25 | 32.00 |

Table 6: Effect of dropout on teacher models. Scores (validation set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.61 | 34.12 | 33.63 | 27.04 | 23.23 | 31.79 | 29.40 |
| Gold + Pseudo (Cross-Ensemble) | 27.10 | 33.79 | 33.17 | 27.43 | 22.78 | 30.66 | 29.16 |
| Gold + Pseudo (Mono-Ensemble) | 29.74 | 36.06 | 36.26 | 28.79 | 27.94 | 33.27 | 32.01 |

Table 7: Choice of teacher models. Scores (validation set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.61 | 34.12 | 33.63 | 27.04 | 23.23 | 31.79 | 29.40 |
| Gold + Pseudo (High) | 29.07 | 36.05 | 35.65 | 30.00 | 26.80 | 32.04 | 31.60 |
| Gold + Pseudo (Low) | 29.27 | 35.51 | 35.76 | 29.30 | 26.84 | 32.20 | 31.48 |
| Gold + Pseudo (Diverse) | 30.38 | 36.27 | 37.11 | 29.71 | 27.59 | 32.75 | 32.30 |

Table 8: Characteristics of pseudo-references. Scores (validation set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.61 | 34.12 | 33.63 | 27.04 | 23.23 | 31.79 | 29.40 |
| Gold + Pseudo-1 | 28.26 | 35.62 | 34.60 | 28.92 | 27.01 | 31.32 | 30.96 |
| Gold + Pseudo-32 | 29.26 | 35.43 | 35.72 | 28.86 | 27.95 | 32.37 | 31.60 |
| Gold + Pseudo-96 | 29.76 | 36.02 | 36.23 | 28.86 | 27.90 | 33.25 | 32.00 |

Table 9: Varying the size of the pseudo-reference pool. Scores (validation set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 26.61 | 34.12 | 33.63 | 27.04 | 23.23 | 31.79 | 29.40 |
| Gold (Soft) | 29.40 | 35.69 | 36.11 | 27.75 | 24.87 | 30.14 | 30.66 |
| Gold (One-Hot) + Pseudo (Soft) | 29.74 | 36.06 | 36.26 | 28.79 | 27.94 | 33.27 | 32.01 |

Table 10: Labeling the gold summary with teacher models. Scores (validation set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.61 | 34.12 | 33.63 | 27.04 | 23.23 | 31.79 | 29.40 |
| Pseudo | 30.26 | 36.11 | 36.82 | 29.02 | 26.82 | 31.69 | 31.79 |
| Gold + Pseudo | 29.74 | 36.06 | 36.26 | 28.79 | 27.94 | 33.27 | 32.01 |

Table 11: Training with only pseudo-labels. Scores (validation set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.16 | 34.04 | 33.99 | 26.89 | 23.44 | 30.40 | 29.15 |
| Gold + Pseudo (Static) | 29.24 | 36.33 | 36.16 | 29.41 | 26.44 | 32.30 | 31.65 |
| Gold + Pseudo (Dropout) | 29.29 | 36.67 | 36.51 | 28.87 | 27.56 | 31.66 | 31.76 |

Table 12: Effect of dropout on teacher models. Scores (test set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.16 | 34.04 | 33.99 | 26.89 | 23.44 | 30.40 | 29.15 |
| Gold + Pseudo (Cross-Ensemble) | 26.20 | 34.33 | 33.53 | 26.51 | 23.45 | 29.43 | 28.91 |
| Gold + Pseudo (Mono-Ensemble) | 29.29 | 36.67 | 36.51 | 28.87 | 27.56 | 31.66 | 31.76 |

Table 13: Choice of teacher models. Scores (test set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.16 | 34.04 | 33.99 | 26.89 | 23.44 | 30.40 | 29.15 |
| Gold + Pseudo (High) | 28.65 | 36.11 | 36.26 | 28.80 | 26.53 | 31.31 | 31.28 |
| Gold + Pseudo (Low) | 29.19 | 35.91 | 36.25 | 28.17 | 26.91 | 31.27 | 31.28 |
| Gold + Pseudo (Diverse) | 29.95 | 36.77 | 37.56 | 28.67 | 27.05 | 32.44 | 32.07 |

Table 14: Characteristics of pseudo-references. Scores (test set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.16 | 34.04 | 33.99 | 26.89 | 23.44 | 30.40 | 29.15 |
| Gold + Pseudo-1 | 28.39 | 35.45 | 35.50 | 28.34 | 26.08 | 30.50 | 30.71 |
| Gold + Pseudo-32 | 29.46 | 35.98 | 36.33 | 28.84 | 26.46 | 31.69 | 31.46 |
| Gold + Pseudo-96 | 29.29 | 36.67 | 36.51 | 28.87 | 27.56 | 31.66 | 31.76 |

Table 15: Varying the size of the pseudo-reference pool. Scores (test set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 26.16 | 34.04 | 33.99 | 26.89 | 23.44 | 30.40 | 29.15 |
| Gold (Soft) | 29.02 | 35.94 | 36.36 | 26.98 | 24.60 | 29.09 | 30.33 |
| Gold (One-Hot) + Pseudo (Soft) | 29.29 | 36.67 | 36.51 | 28.87 | 27.56 | 31.66 | 31.76 |

Table 16: Labeling the gold summary with teacher models. Scores (test set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 26.16 | 34.04 | 33.99 | 26.89 | 23.44 | 30.40 | 29.15 |
| Pseudo | 29.86 | 36.67 | 37.55 | 28.26 | 25.47 | 30.97 | 31.46 |
| Gold + Pseudo | 29.29 | 36.67 | 36.51 | 28.87 | 27.56 | 31.66 | 31.76 |

Table 17: Training with only pseudo-labels. Scores (test set) are displayed in ROUGE-1.

| | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trans-Sum | 8.92 | 13.70 | 9.31 | 19.34 | 29.10 | 19.87 | 12.54 | 12.17 | 11.34 | 11.97 | 16.12 | 19.58 | 16.53 | 28.96 | 16.51 | 13.87 | 16.97 | 12.10 | 14.04 | 11.97 | 16.10 | 15.76 |
| Sum-Trans | 22.27 | 23.15 | 26.85 | 28.28 | 37.80 | 32.63 | 23.14 | 23.68 | 25.28 | 28.76 | 30.11 | 34.27 | 29.49 | 39.55 | 31.73 | 32.27 | 30.68 | 27.19 | 28.18 | 26.53 | 30.38 | 29.15 |
| Gold | 23.44 | 30.40 | 26.89 | 36.45 | 39.35 | 34.04 | 26.21 | 25.74 | 26.16 | 29.60 | 33.99 | 35.41 | 34.11 | 40.93 | 33.29 | 33.75 | 32.75 | 28.63 | 32.54 | 28.25 | 33.40 | 31.68 |
| Gold + OTSum | 25.54 | 29.64 | 28.87 | 35.03 | 42.05 | 35.48 | 28.11 | 26.09 | 27.61 | 30.52 | 31.04 | 37.13 | 32.37 | 42.35 | 32.55 | 33.45 | 35.82 | 27.31 | 30.15 | 27.76 | 32.42 | 31.97 |
| Gold + Many2Many | 25.47 | 33.81 | 26.64 | 39.33 | 41.47 | 34.92 | 27.48 | 26.69 | 27.06 | 29.80 | 35.35 | 35.07 | 35.36 | 41.48 | 34.40 | 34.53 | 35.39 | 27.78 | 33.21 | 27.25 | 33.39 | 32.66 |
| Gold + Pseudo | 27.56 | 31.66 | 28.87 | 39.24 | 41.85 | 36.67 | 28.81 | 28.25 | 29.29 | 32.21 | 36.51 | 37.26 | 36.29 | 42.35 | 35.72 | 34.47 | 35.97 | 30.23 | 35.13 | 29.05 | 35.73 | 33.96 |

Table 18: Fully Supervised Training with mBART-50. Scores (test set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Avg. |
|---|---|---|---|
| [Transformer] Gold | 20.64 | 28.61 | 24.62 |
| [Transformer] Gold + Pseudo | 21.84 | 29.01 | 25.43 |
| [PISCES] Gold | 28.61 | 36.29 | 32.45 |
| [PISCES] Gold + Pseudo | 30.33 | 37.11 | 33.72 |

Table 19: Fully Supervised Training with other architectures. Scores (test set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [50] Gold | 16.74 | 23.88 | 20.31 | 13.25 | 12.81 | 17.75 | 17.46 |
| [50] Gold + Pseudo | 17.29 | 26.33 | 23.48 | 15.46 | 14.29 | 19.54 | 19.40 |
| [100] Gold | 17.51 | 25.39 | 21.94 | 14.96 | 13.93 | 19.65 | 18.90 |
| [100] Gold + Pseudo | 19.10 | 27.95 | 24.42 | 16.46 | 16.29 | 21.05 | 20.88 |
| [300] Gold | 18.39 | 27.66 | 24.49 | 17.48 | 17.17 | 21.41 | 21.10 |
| [300] Gold + Pseudo | 21.86 | 30.30 | 25.85 | 20.30 | 20.06 | 25.53 | 23.98 |

Table 20: Few-Shot Learning (50/100/300-shot). Scores (test set) are displayed in ROUGE-1.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [LoRA] Gold | 23.67 | 32.52 | 31.78 | 23.85 | 20.37 | 28.27 | 26.74 |
| [LoRA] Gold + Pseudo | 26.26 | 34.26 | 33.44 | 24.72 | 21.43 | 29.76 | 28.31 |

Table 21: Low-Rank Adaptation. Scores (test set) are displayed in ROUGE-1.

### D.2 ROUGE-2

| | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mBART-50 | 8.63 | 10.01 | 5.25 | 8.00 | 8.37 | 9.17 | 6.08 | 7.51 | 9.62 | 8.66 | 7.72 | 8.23 | 6.92 | 6.56 | 6.64 | 5.98 | 5.87 | 6.75 | 5.76 | 3.31 | 5.80 |
| Transformer-Large | 4.27 | 2.09 | 1.87 | 1.64 | 2.82 | 3.58 | 0.58 | 1.52 | 4.10 | 4.34 | 1.25 | 3.95 | 1.35 | 3.58 | 1.03 | 2.49 | 1.03 | 2.08 | 1.88 | 1.23 | 3.60 |
| Transformer-Base | 0.91 | 2.03 | 2.68 | 2.60 | 2.03 | 3.22 | 1.90 | 1.96 | 3.97 | 4.38 | 1.44 | 3.83 | 1.69 | 3.42 | 1.62 | 2.97 | 1.20 | 2.76 | 2.28 | 2.42 | 3.68 |
| Transformer-Small | 2.97 | 3.57 | 0.85 | 5.52 | 1.16 | 3.14 | 1.72 | 1.51 | 3.32 | 3.14 | 1.12 | 3.65 | 1.15 | 3.17 | 0.87 | 2.94 | 2.57 | 2.50 | 1.64 | 1.70 | 2.88 |

Table 22: Quantifying the cross-/mono-lingual summarization gap. Scores (validation set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 10.02 | 14.31 | 13.63 | 13.65 | 11.07 | 15.67 | 13.06 |
| Gold + Pseudo (Static) | 12.40 | 16.30 | 15.52 | 15.55 | 13.78 | 17.14 | 15.12 |
| Gold + Pseudo (Dropout) | 12.79 | 16.21 | 15.77 | 14.75 | 14.12 | 17.02 | 15.11 |

Table 23: Effect of dropout on teacher models. Scores (validation set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 10.02 | 14.31 | 13.63 | 13.65 | 11.07 | 15.67 | 13.06 |
| Gold + Pseudo (Cross-Ensemble) | 10.45 | 14.32 | 13.34 | 13.61 | 10.12 | 14.86 | 12.78 |
| Gold + Pseudo (Mono-Ensemble) | 12.78 | 16.28 | 15.78 | 14.70 | 14.13 | 17.00 | 15.11 |

Table 24: Choice of teacher models. Scores (validation set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 10.02 | 14.31 | 13.63 | 13.65 | 11.07 | 15.67 | 13.06 |
| Gold + Pseudo (High) | 12.42 | 16.33 | 15.49 | 15.53 | 13.45 | 16.16 | 14.90 |
| Gold + Pseudo (Low) | 12.42 | 15.83 | 15.28 | 15.34 | 13.12 | 16.26 | 14.71 |
| Gold + Pseudo (Diverse) | 13.58 | 17.21 | 16.64 | 15.35 | 14.08 | 17.07 | 15.66 |

Table 25: Characteristics of pseudo-references. Scores (validation set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 10.02 | 14.31 | 13.63 | 13.65 | 11.07 | 15.67 | 13.06 |
| Gold + Pseudo-1 | 11.59 | 15.42 | 14.47 | 14.67 | 13.52 | 15.64 | 14.22 |
| Gold + Pseudo-32 | 12.50 | 15.61 | 15.35 | 14.81 | 14.18 | 16.35 | 14.80 |
| Gold + Pseudo-96 | 12.79 | 16.21 | 15.77 | 14.75 | 14.12 | 17.02 | 15.11 |

Table 26: Varying the size of the pseudo-reference pool. Scores (validation set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 10.02 | 14.31 | 13.63 | 13.65 | 11.07 | 15.67 | 13.06 |
| Gold (Soft) | 12.88 | 16.57 | 15.82 | 14.06 | 11.34 | 14.96 | 14.27 |
| Gold (One-Hot) + Pseudo (Soft) | 12.78 | 16.28 | 15.78 | 14.70 | 14.13 | 17.00 | 15.11 |

Table 27: Labeling the gold summary with teacher models. Scores (validation set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 10.02 | 14.31 | 13.63 | 13.65 | 11.07 | 15.67 | 13.06 |
| Pseudo | 13.50 | 16.51 | 16.38 | 14.82 | 12.69 | 16.08 | 15.00 |
| Gold + Pseudo | 12.78 | 16.28 | 15.78 | 14.70 | 14.13 | 17.00 | 15.11 |

Table 28: Training with only pseudo-labels. Scores (validation set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 9.96 | 14.11 | 14.21 | 13.37 | 10.53 | 14.67 | 12.81 |
| Gold + Pseudo (Static) | 12.42 | 16.72 | 15.74 | 15.09 | 12.81 | 16.46 | 14.87 |
| Gold + Pseudo (Dropout) | 12.69 | 16.97 | 16.06 | 14.90 | 13.82 | 16.06 | 15.08 |

Table 29: Effect of dropout on teacher models. Scores (test set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 9.96 | 14.11 | 14.21 | 13.37 | 10.53 | 14.67 | 12.81 |
| Gold + Pseudo (Cross-Ensemble) | 9.92 | 14.63 | 13.64 | 12.92 | 10.68 | 14.07 | 12.64 |
| Gold + Pseudo (Mono-Ensemble) | 12.69 | 16.97 | 16.06 | 14.90 | 13.82 | 16.06 | 15.08 |

Table 30: Choice of teacher models. Scores (test set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 9.96 | 14.11 | 14.21 | 13.37 | 10.53 | 14.67 | 12.81 |
| Gold + Pseudo (High) | 12.13 | 16.24 | 15.96 | 14.61 | 12.68 | 15.68 | 14.55 |
| Gold + Pseudo (Low) | 12.50 | 16.17 | 16.00 | 14.22 | 12.89 | 15.68 | 14.58 |
| Gold + Pseudo (Diverse) | 13.38 | 17.32 | 16.96 | 14.58 | 12.95 | 16.98 | 15.36 |

Table 31: Characteristics of pseudo-references. Scores (test set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 9.96 | 14.11 | 14.21 | 13.37 | 10.53 | 14.67 | 12.81 |
| Gold + Pseudo-1 | 11.44 | 15.26 | 15.03 | 14.20 | 12.20 | 14.79 | 13.82 |
| Gold + Pseudo-32 | 12.57 | 16.27 | 16.02 | 14.73 | 12.84 | 16.04 | 14.74 |
| Gold + Pseudo-96 | 12.69 | 16.97 | 16.06 | 14.90 | 13.82 | 16.06 | 15.08 |

Table 32: Varying the size of the pseudo-reference pool. Scores (test set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 9.96 | 14.11 | 14.21 | 13.37 | 10.53 | 14.67 | 12.81 |
| Gold (Soft) | 12.63 | 16.84 | 15.98 | 13.46 | 11.70 | 14.59 | 14.20 |
| Gold (One-Hot) + Pseudo (Soft) | 12.69 | 16.97 | 16.06 | 14.90 | 13.82 | 16.06 | 15.08 |

Table 33: Labeling the gold summary with teacher models. Scores (test set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 9.96 | 14.11 | 14.21 | 13.37 | 10.53 | 14.67 | 12.81 |
| Pseudo | 13.32 | 17.36 | 16.83 | 14.31 | 12.21 | 15.79 | 14.97 |
| Gold + Pseudo | 12.69 | 16.97 | 16.06 | 14.90 | 13.82 | 16.06 | 15.08 |

Table 34: Training with only pseudo-labels. Scores (test set) are displayed in ROUGE-2.

| | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trans-Sum | 2.58 | 4.21 | 2.69 | 3.81 | 10.18 | 3.50 | 2.86 | 1.88 | 1.36 | 1.85 | 2.73 | 3.10 | 2.95 | 10.33 | 2.80 | 3.74 | 3.17 | 1.96 | 1.98 | 2.59 | 2.67 | 3.47 |
| Sum-Trans | 8.84 | 9.85 | 12.20 | 9.78 | 17.28 | 12.73 | 9.53 | 8.67 | 9.09 | 11.89 | 10.84 | 14.22 | 10.33 | 18.09 | 11.80 | 15.58 | 11.44 | 11.38 | 9.26 | 11.85 | 11.05 | 11.70 |
| Gold | 10.53 | 14.67 | 13.37 | 16.52 | 18.47 | 14.11 | 11.78 | 9.83 | 9.96 | 12.73 | 14.21 | 15.24 | 13.96 | 19.70 | 13.08 | 16.63 | 12.84 | 12.46 | 12.56 | 13.11 | 13.13 | 13.76 |
| Gold + OTSum | 12.07 | 13.90 | 14.85 | 14.79 | 20.93 | 15.60 | 13.44 | 10.60 | 10.94 | 13.55 | 11.58 | 16.99 | 11.81 | 20.79 | 12.41 | 16.58 | 15.30 | 11.34 | 10.75 | 12.97 | 12.70 | 13.99 |
| Gold + Many2Many | 11.22 | 17.10 | 12.40 | 18.53 | 20.41 | 14.76 | 13.17 | 10.97 | 10.57 | 12.75 | 15.06 | 15.02 | 14.92 | 20.04 | 13.91 | 17.32 | 15.07 | 11.73 | 13.35 | 12.73 | 13.46 | 14.50 |
| Gold + Pseudo | 13.82 | 16.06 | 14.90 | 18.79 | 20.98 | 16.97 | 14.32 | 12.57 | 12.69 | 15.24 | 16.06 | 17.69 | 15.94 | 21.14 | 15.49 | 17.77 | 15.70 | 14.39 | 15.02 | 14.65 | 15.56 | 15.99 |

Table 35: Fully Supervised Training with mBART-50. Scores (test set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Avg. |
|---|---|---|---|
| [Transformer] Gold | 6.37 | 10.29 | 8.33 |
| [Transformer] Gold + Pseudo | 7.76 | 11.41 | 9.59 |
| [PISCES] Gold | 11.46 | 15.82 | 13.64 |
| [PISCES] Gold + Pseudo | 13.35 | 17.64 | 15.50 |

Table 36: Fully Supervised Training with other architectures. Scores (test set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [50] Gold | 3.50 | 5.71 | 4.17 | 3.80 | 3.26 | 6.35 | 4.46 |
| [50] Gold + Pseudo | 3.64 | 8.21 | 5.91 | 5.19 | 4.82 | 7.16 | 5.82 |
| [100] Gold | 3.64 | 6.55 | 5.13 | 4.73 | 4.07 | 6.82 | 5.16 |
| [100] Gold + Pseudo | 4.67 | 8.72 | 6.65 | 5.72 | 5.60 | 8.20 | 6.59 |
| [300] Gold | 4.02 | 8.87 | 6.73 | 6.00 | 5.41 | 8.02 | 6.51 |
| [300] Gold + Pseudo | 6.25 | 10.90 | 7.71 | 8.18 | 7.97 | 10.94 | 8.66 |

Table 37: Few-Shot Learning (50/100/300-shot). Scores (test set) are displayed in ROUGE-2.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [LoRA] Gold | 7.78 | 12.62 | 11.81 | 9.92 | 7.16 | 12.42 | 10.29 |
| [LoRA] Gold + Pseudo | 10.17 | 14.35 | 13.50 | 11.61 | 8.95 | 14.55 | 12.19 |

Table 38: Low-Rank Adaptation. Scores (test set) are displayed in ROUGE-2.

### D.3 ROUGE-L

| | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mBART-50 | 8.78 | 9.67 | 5.01 | 8.28 | 6.52 | 6.88 | 6.09 | 6.64 | 9.13 | 7.98 | 7.36 | 6.15 | 6.41 | 5.26 | 6.49 | 5.33 | 5.85 | 6.05 | 5.01 | 2.82 | 5.56 |
| Transformer-Large | 5.19 | 2.40 | 2.68 | 1.92 | 2.88 | 3.02 | 1.10 | 1.46 | 4.76 | 4.92 | 1.34 | 3.49 | 1.67 | 3.40 | 1.24 | 2.95 | 1.29 | 2.48 | 2.46 | 1.56 | 3.76 |
| Transformer-Base | 1.79 | 2.20 | 3.44 | 2.56 | 1.92 | 3.16 | 2.12 | 2.30 | 4.40 | 4.93 | 1.42 | 3.34 | 2.20 | 3.25 | 2.02 | 3.45 | 1.06 | 2.95 | 2.67 | 2.63 | 3.97 |
| Transformer-Small | 3.77 | 4.53 | 1.34 | 5.73 | 1.50 | 3.19 | 2.10 | 2.37 | 3.97 | 3.53 | 1.38 | 3.30 | 1.61 | 2.89 | 1.20 | 3.57 | 2.79 | 2.87 | 2.00 | 1.94 | 3.15 |

Table 39: Quantifying the cross-/mono-lingual summarization gap. Scores (validation set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.38 | 27.26 | 26.64 | 22.17 | 18.83 | 24.88 | 23.36 |
| Gold + Pseudo (Static) | 22.81 | 29.18 | 28.53 | 24.57 | 22.36 | 26.60 | 25.68 |
| Gold + Pseudo (Dropout) | 23.39 | 28.87 | 28.77 | 23.65 | 22.78 | 26.42 | 25.65 |

Table 40: Effect of dropout on teacher models. Scores (validation set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.38 | 27.26 | 26.64 | 22.17 | 18.83 | 24.88 | 23.36 |
| Gold + Pseudo (Cross-Ensemble) | 20.98 | 27.22 | 26.25 | 22.60 | 18.68 | 23.80 | 23.25 |
| Gold + Pseudo (Mono-Ensemble) | 23.39 | 28.91 | 28.78 | 23.60 | 22.72 | 26.43 | 25.64 |

Table 41: Choice of teacher models. Scores (validation set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.38 | 27.26 | 26.64 | 22.17 | 18.83 | 24.88 | 23.36 |
| Gold + Pseudo (High) | 22.87 | 29.24 | 28.37 | 24.38 | 21.82 | 25.17 | 25.31 |
| Gold + Pseudo (Low) | 22.98 | 28.82 | 28.35 | 24.08 | 21.68 | 25.58 | 25.25 |
| Gold + Pseudo (Diverse) | 23.95 | 29.67 | 29.59 | 24.28 | 22.21 | 26.76 | 26.08 |

Table 42: Characteristics of pseudo-references. Scores (validation set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.38 | 27.26 | 26.64 | 22.17 | 18.83 | 24.88 | 23.36 |
| Gold + Pseudo-1 | 21.95 | 28.28 | 27.50 | 23.57 | 22.03 | 24.96 | 24.71 |
| Gold + Pseudo-32 | 23.00 | 28.59 | 28.34 | 23.43 | 22.59 | 25.99 | 25.32 |
| Gold + Pseudo-96 | 23.39 | 28.87 | 28.77 | 23.65 | 22.78 | 26.42 | 25.65 |

Table 43: Varying the size of the pseudo-reference pool. Scores (validation set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 20.38 | 27.26 | 26.64 | 22.17 | 18.83 | 24.88 | 23.36 |
| Gold (Soft) | 23.47 | 29.41 | 28.68 | 22.78 | 19.81 | 24.09 | 24.71 |
| Gold (One-Hot) + Pseudo (Soft) | 23.39 | 28.91 | 28.78 | 23.60 | 22.72 | 26.43 | 25.64 |

Table 44: Labeling the gold summary with teacher models. Scores (validation set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.38 | 27.26 | 26.64 | 22.17 | 18.83 | 24.88 | 23.36 |
| Pseudo | 23.92 | 29.61 | 29.34 | 23.54 | 21.23 | 25.59 | 25.54 |
| Gold + Pseudo | 23.39 | 28.91 | 28.78 | 23.60 | 22.72 | 26.43 | 25.64 |

Table 45: Training with only pseudo-labels. Scores (validation set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.27 | 27.26 | 26.97 | 21.89 | 18.83 | 23.75 | 23.16 |
| Gold + Pseudo (Static) | 22.91 | 29.66 | 28.56 | 23.80 | 21.34 | 25.18 | 25.24 |
| Gold + Pseudo (Dropout) | 23.00 | 29.91 | 28.95 | 23.51 | 22.31 | 25.09 | 25.46 |

Table 46: Effect of dropout on teacher models. Scores (test set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.27 | 27.26 | 26.97 | 21.89 | 18.83 | 23.75 | 23.16 |
| Gold + Pseudo (Cross-Ensemble) | 20.36 | 27.95 | 26.46 | 21.42 | 18.62 | 22.68 | 22.91 |
| Gold + Pseudo (Mono-Ensemble) | 23.00 | 29.91 | 28.95 | 23.51 | 22.31 | 25.09 | 25.46 |

Table 47: Choice of teacher models. Scores (test set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.27 | 27.26 | 26.97 | 21.89 | 18.83 | 23.75 | 23.16 |
| Gold + Pseudo (High) | 22.61 | 29.32 | 28.82 | 23.33 | 21.20 | 24.31 | 24.93 |
| Gold + Pseudo (Low) | 22.83 | 29.19 | 28.91 | 23.04 | 21.46 | 24.65 | 25.01 |
| Gold + Pseudo (Diverse) | 23.76 | 30.28 | 29.90 | 23.37 | 21.47 | 25.91 | 25.78 |

Table 48: Characteristics of pseudo-references. Scores (test set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.27 | 27.26 | 26.97 | 21.89 | 18.83 | 23.75 | 23.16 |
| Gold + Pseudo-1 | 21.95 | 28.23 | 28.03 | 22.81 | 20.99 | 23.69 | 24.28 |
| Gold + Pseudo-32 | 23.09 | 29.08 | 28.94 | 23.39 | 21.37 | 24.98 | 25.14 |
| Gold + Pseudo-96 | 23.00 | 29.91 | 28.95 | 23.51 | 22.31 | 25.09 | 25.46 |

Table 49: Varying the size of the pseudo-reference pool. Scores (test set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 20.27 | 27.26 | 26.97 | 21.89 | 18.83 | 23.75 | 23.16 |
| Gold (Soft) | 23.03 | 29.87 | 28.89 | 21.88 | 19.70 | 23.27 | 24.44 |
| Gold (One-Hot) + Pseudo (Soft) | 23.00 | 29.91 | 28.95 | 23.51 | 22.31 | 25.09 | 25.46 |

Table 50: Labeling the gold summary with teacher models. Scores (test set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 20.27 | 27.26 | 26.97 | 21.89 | 18.83 | 23.75 | 23.16 |
| Pseudo | 23.58 | 30.22 | 29.76 | 22.58 | 20.66 | 24.60 | 25.23 |
| Gold + Pseudo | 23.00 | 29.91 | 28.95 | 23.51 | 22.31 | 25.09 | 25.46 |

Table 51: Training with only pseudo-labels. Scores (test set) are displayed in ROUGE-L.

| | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trans-Sum | 7.21 | 11.04 | 7.42 | 14.73 | 26.37 | 16.64 | 9.43 | 9.78 | 9.64 | 10.09 | 12.55 | 16.63 | 13.02 | 26.28 | 12.90 | 11.05 | 13.11 | 9.59 | 11.17 | 9.01 | 12.59 | 12.87 |
| Sum-Trans | 17.48 | 17.84 | 21.76 | 21.35 | 32.68 | 26.34 | 17.54 | 17.60 | 19.51 | 22.65 | 23.23 | 27.84 | 22.78 | 33.80 | 24.79 | 25.86 | 23.85 | 20.62 | 21.48 | 20.27 | 23.48 | 22.99 |
| Gold | 18.83 | 23.75 | 21.89 | 29.18 | 33.43 | 27.26 | 19.97 | 18.89 | 20.27 | 23.51 | 26.97 | 28.54 | 26.92 | 35.33 | 25.93 | 26.91 | 25.31 | 21.79 | 25.12 | 21.56 | 25.98 | 25.11 |
| Gold + OTSum | 20.39 | 22.70 | 23.18 | 27.71 | 36.10 | 28.92 | 21.79 | 19.57 | 21.42 | 24.39 | 23.97 | 30.29 | 24.55 | 36.66 | 25.08 | 26.37 | 28.20 | 20.59 | 23.08 | 21.39 | 25.17 | 25.31 |
| Gold + Many2Many | 20.55 | 26.91 | 21.79 | 32.25 | 35.97 | 28.48 | 21.55 | 20.35 | 21.34 | 23.67 | 28.55 | 28.78 | 28.52 | 35.97 | 27.45 | 28.03 | 28.47 | 21.25 | 26.46 | 21.45 | 26.63 | 26.40 |
| Gold + Pseudo | 22.31 | 25.09 | 23.51 | 31.75 | 35.73 | 29.91 | 22.66 | 21.53 | 23.00 | 25.57 | 28.95 | 30.65 | 28.96 | 36.47 | 28.24 | 27.79 | 28.51 | 23.54 | 27.78 | 23.03 | 28.46 | 27.31 |

Table 52: Fully Supervised Training with mBART-50. Scores (test set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Avg. |
|---|---|---|---|
| [Transformer] Gold | 15.33 | 22.09 | 18.71 |
| [Transformer] Gold + Pseudo | 16.43 | 23.16 | 19.80 |
| [PISCES] Gold | 22.18 | 29.25 | 25.71 |
| [PISCES] Gold + Pseudo | 23.95 | 30.54 | 27.24 |

Table 53: Fully Supervised Training with other architectures. Scores (test set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [50] Gold | 12.84 | 18.55 | 14.85 | 9.77 | 9.61 | 13.05 | 13.11 |
| [50] Gold + Pseudo | 12.93 | 21.22 | 16.92 | 11.76 | 11.04 | 14.70 | 14.76 |
| [100] Gold | 13.46 | 19.94 | 16.07 | 11.16 | 10.44 | 14.47 | 14.26 |
| [100] Gold + Pseudo | 14.38 | 22.29 | 18.09 | 12.62 | 12.80 | 15.81 | 16.00 |
| [300] Gold | 13.67 | 22.04 | 18.07 | 12.97 | 12.73 | 16.06 | 15.92 |
| [300] Gold + Pseudo | 16.46 | 24.25 | 19.35 | 16.00 | 15.59 | 19.04 | 18.45 |

Table 54: Few-Shot Learning (50/100/300-shot). Scores (test set) are displayed in ROUGE-L.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [LoRA] Gold | 18.26 | 26.39 | 24.44 | 18.04 | 14.97 | 20.84 | 20.49 |
| [LoRA] Gold + Pseudo | 20.68 | 27.85 | 26.07 | 19.68 | 16.85 | 23.26 | 22.40 |

Table 55: Low-Rank Adaptation. Scores (test set) are displayed in ROUGE-L.

### D.4 SacreBLEU (Post, 2018)

| | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mBART-50 | 0.58 | 5.26 | 0.73 | 4.83 | 6.44 | 3.41 | 1.87 | 3.45 | 1.57 | 2.13 | 5.31 | 5.06 | 3.04 | 5.36 | 4.12 | 2.72 | 3.09 | 3.32 | 3.17 | 1.48 | 3.60 |
| Transformer-Large | 2.76 | 1.34 | 0.99 | 1.55 | 2.33 | 0.94 | 0.21 | 0.29 | 1.47 | 0.60 | 0.50 | 1.18 | 0.18 | 2.51 | 0.43 | 0.41 | 0.71 | 0.87 | 0.77 | 0.29 | 1.48 |
| Transformer-Base | 0.26 | 1.05 | 2.22 | 2.20 | 1.86 | 0.55 | 1.29 | 0.73 | 2.35 | 0.64 | 0.70 | 0.63 | 0.70 | 1.59 | 0.80 | 0.98 | 0.58 | 0.81 | 1.31 | 0.42 | 1.91 |
| Transformer-Small | 0.82 | 1.98 | 0.56 | 7.03 | 0.68 | 1.54 | 1.59 | 1.06 | 1.61 | -0.84 | 0.53 | 0.73 | 0.31 | 2.34 | 0.26 | 2.46 | 1.90 | 1.07 | 0.73 | 0.13 | 1.07 |

Table 56: Quantifying the cross-/mono-lingual summarization gap. Scores (validation set) are displayed in SacreBLEU.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.35 | 8.89 | 10.88 | 15.06 | 9.38 | 14.39 | 10.66 |
| Gold + Pseudo (Static) | 5.87 | 9.15 | 11.77 | 14.29 | 12.16 | 15.04 | 11.38 |
| Gold + Pseudo (Dropout) | 6.85 | 10.16 | 11.88 | 15.42 | 13.38 | 14.31 | 12.00 |

Table 57: Effect of dropout on teacher models. Scores (validation set) are displayed in SacreBLEU.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.35 | 8.89 | 10.88 | 15.06 | 9.38 | 14.39 | 10.66 |
| Gold + Pseudo (Cross-Ensemble) | 5.59 | 8.68 | 10.02 | 14.04 | 9.33 | 13.81 | 10.24 |
| Gold + Pseudo (Mono-Ensemble) | 6.80 | 10.18 | 11.88 | 15.37 | 13.38 | 14.26 | 11.98 |

Table 58: Choice of teacher models. Scores (validation set) are displayed in SacreBLEU.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.35 | 8.89 | 10.88 | 15.06 | 9.38 | 14.39 | 10.66 |
| Gold + Pseudo (High) | 6.65 | 9.10 | 11.12 | 14.31 | 12.80 | 13.79 | 11.29 |
| Gold + Pseudo (Low) | 6.19 | 9.76 | 10.88 | 14.74 | 12.49 | 13.57 | 11.27 |
| Gold + Pseudo (Diverse) | 6.46 | 9.25 | 12.52 | 14.87 | 12.62 | 12.40 | 11.35 |

Table 59: Characteristics of pseudo-references. Scores (validation set) are displayed in SacreBLEU.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.35 | 8.89 | 10.88 | 15.06 | 9.38 | 14.39 | 10.66 |
| Gold + Pseudo-1 | 6.61 | 9.15 | 11.00 | 14.58 | 12.93 | 14.02 | 11.38 |
| Gold + Pseudo-32 | 6.58 | 9.72 | 11.43 | 15.31 | 13.39 | 13.31 | 11.62 |
| Gold + Pseudo-96 | 6.85 | 10.16 | 11.88 | 15.42 | 13.38 | 14.31 | 12.00 |

Table 60: Varying the size of the pseudo-reference pool. Scores (validation set) are displayed in SacreBLEU.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 5.35 | 8.89 | 10.88 | 15.06 | 9.38 | 14.39 | 10.66 |
| Gold (Soft) | 5.67 | 7.19 | 11.15 | 11.61 | 9.04 | 10.58 | 9.21 |
| Gold (One-Hot) + Pseudo (Soft) | 6.80 | 10.18 | 11.88 | 15.37 | 13.38 | 14.26 | 11.98 |

Table 61: Labeling the gold summary with teacher models. Scores (validation set) are displayed in SacreBLEU.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.35 | 8.89 | 10.88 | 15.06 | 9.38 | 14.39 | 10.66 |
| Pseudo | 5.91 | 8.00 | 11.36 | 12.96 | 10.57 | 11.91 | 10.12 |
| Gold + Pseudo | 6.80 | 10.18 | 11.88 | 15.37 | 13.38 | 14.26 | 11.98 |

Table 62: Training with only pseudo-labels. Scores (validation set) are displayed in SacreBLEU.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.04 | 8.45 | 11.25 | 11.11 | 7.39 | 14.12 | 9.56 |
| Gold + Pseudo (Static) | 5.91 | 9.30 | 11.87 | 11.54 | 9.49 | 14.41 | 10.42 |
| Gold + Pseudo (Dropout) | 6.02 | 10.36 | 11.92 | 12.85 | 10.86 | 13.12 | 10.86 |

Table 63: Effect of dropout on teacher models. Scores (test set) are displayed in SacreBLEU.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.04 | 8.45 | 11.25 | 11.11 | 7.39 | 14.12 | 9.56 |
| Gold + Pseudo (Cross-Ensemble) | 4.83 | 8.69 | 10.32 | 11.20 | 8.94 | 12.82 | 9.47 |
| Gold + Pseudo (Mono-Ensemble) | 6.02 | 10.36 | 11.92 | 12.85 | 10.86 | 13.12 | 10.86 |

Table 64: Choice of teacher models. Scores (test set) are displayed in SacreBLEU.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.04 | 8.45 | 11.25 | 11.11 | 7.39 | 14.12 | 9.56 |
| Gold + Pseudo (High) | 5.53 | 9.43 | 11.14 | 11.04 | 9.22 | 13.62 | 10.00 |
| Gold + Pseudo (Low) | 5.48 | 9.62 | 11.19 | 11.75 | 9.60 | 13.67 | 10.22 |
| Gold + Pseudo (Diverse) | 6.19 | 9.45 | 12.55 | 12.46 | 9.53 | 13.11 | 10.55 |

Table 65: Characteristics of pseudo-references. Scores (test set) are displayed in SacreBLEU.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.04 | 8.45 | 11.25 | 11.11 | 7.39 | 14.12 | 9.56 |
| Gold + Pseudo-1 | 5.52 | 9.61 | 11.12 | 11.03 | 9.90 | 13.40 | 10.10 |
| Gold + Pseudo-32 | 5.77 | 9.52 | 11.92 | 11.65 | 9.67 | 13.97 | 10.42 |
| Gold + Pseudo-96 | 6.02 | 10.36 | 11.92 | 12.85 | 10.86 | 13.12 | 10.86 |

Table 66: Varying the size of the pseudo-reference pool. Scores (test set) are displayed in SacreBLEU.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 5.04 | 8.45 | 11.25 | 11.11 | 7.39 | 14.12 | 9.56 |
| Gold (Soft) | 4.59 | 7.71 | 11.17 | 9.91 | 9.33 | 10.74 | 8.91 |
| Gold (One-Hot) + Pseudo (Soft) | 6.02 | 10.36 | 11.92 | 12.85 | 10.86 | 13.12 | 10.86 |

Table 67: Labeling the gold summary with teacher models. Scores (test set) are displayed in SacreBLEU.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 5.04 | 8.45 | 11.25 | 11.11 | 7.39 | 14.12 | 9.56 |
| Pseudo | 4.81 | 8.48 | 11.56 | 11.30 | 9.14 | 11.51 | 9.47 |
| Gold + Pseudo | 6.02 | 10.36 | 11.92 | 12.85 | 10.86 | 13.12 | 10.86 |

Table 68: Training with only pseudo-labels. Scores (test set) are displayed in SacreBLEU.

| | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trans-Sum | 6.23 | 3.94 | 6.12 | 3.03 | 2.96 | 1.50 | 3.20 | 0.81 | 0.48 | 1.64 | 2.02 | 1.27 | 2.15 | 2.59 | 2.12 | 2.88 | 2.37 | 1.31 | 1.35 | 2.63 | 1.85 | 2.50 |
| Sum-Trans | 7.54 | 9.13 | 8.27 | 6.73 | 9.35 | 6.75 | 6.45 | 4.89 | 4.23 | 5.78 | 6.95 | 7.02 | 6.64 | 10.74 | 7.53 | 11.36 | 7.51 | 6.55 | 6.29 | 6.64 | 6.74 | 7.29 |
| Gold | 7.39 | 14.12 | 11.11 | 15.08 | 12.35 | 8.45 | 9.11 | 6.81 | 5.04 | 7.64 | 11.25 | 8.76 | 10.86 | 10.17 | 9.95 | 12.99 | 9.64 | 7.45 | 9.24 | 8.13 | 9.24 | 9.75 |
| Gold + OTSum | 9.89 | 11.40 | 10.13 | 11.82 | 11.94 | 8.18 | 9.24 | 5.16 | 4.24 | 6.54 | 8.41 | 8.86 | 9.12 | 11.30 | 8.61 | 11.76 | 11.18 | 6.26 | 7.89 | 7.78 | 8.35 | 8.96 |
| Gold + Many2Many | 9.76 | 14.36 | 12.00 | 14.42 | 12.91 | 9.14 | 9.46 | 7.40 | 5.27 | 8.29 | 11.24 | 9.11 | 11.11 | 11.91 | 9.94 | 13.15 | 10.95 | 7.29 | 9.49 | 7.79 | 9.11 | 10.20 |
| Gold + Pseudo | 10.86 | 13.12 | 12.85 | 15.40 | 13.07 | 10.36 | 9.83 | 7.49 | 6.02 | 8.56 | 11.92 | 9.36 | 11.62 | 11.94 | 10.79 | 12.67 | 11.37 | 7.21 | 10.12 | 7.76 | 9.97 | 10.59 |

Table 69: Fully Supervised Training with mBART-50. Scores (test set) are displayed in SacreBLEU.

| | Ar2Ko | Ko2Ja | Avg. |
|---|---|---|---|
| [Transformer] Gold | 2.91 | 8.16 | 5.54 |
| [Transformer] Gold + Pseudo | 4.30 | 6.88 | 5.59 |
| [PISCES] Gold | 5.58 | 9.57 | 7.58 |
| [PISCES] Gold + Pseudo | 5.94 | 9.39 | 7.67 |

Table 70: Fully Supervised Training with other architectures. Scores (test set) are displayed in SacreBLEU.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [50] Gold | 0.59 | 3.44 | 2.79 | 7.00 | 2.48 | 3.63 | 3.32 |
| [50] Gold + Pseudo | 2.01 | 3.36 | 4.65 | 8.62 | 8.36 | 6.54 | 5.59 |
| [100] Gold | 0.43 | 3.91 | 3.45 | 6.83 | 3.09 | 5.56 | 3.88 |
| [100] Gold + Pseudo | 2.98 | 4.89 | 4.56 | 8.11 | 8.23 | 8.06 | 6.14 |
| [300] Gold | 2.39 | 5.11 | 4.74 | 8.44 | 6.03 | 6.78 | 5.58 |
| [300] Gold + Pseudo | 3.24 | 6.96 | 5.77 | 8.42 | 7.24 | 9.29 | 6.82 |

Table 71: Few-Shot Learning (50/100/300-shot). Scores (test set) are displayed in SacreBLEU.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [LoRA] Gold | 2.78 | 7.62 | 8.80 | 10.28 | 6.17 | 12.05 | 7.95 |
| [LoRA] Gold + Pseudo | 4.01 | 7.35 | 9.42 | 11.71 | 7.71 | 11.49 | 8.62 |

Table 72: Low-Rank Adaptation. Scores (test set) are displayed in SacreBLEU.

### D.5 BERTScore (Zhang et al., 2020)

Results were computed with the *bert-base-multilingual-cased* model[14].

| | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mBART-50 | 3.00 | 2.74 | 0.64 | 2.59 | 2.04 | 2.30 | 1.99 | 1.92 | 2.94 | 2.65 | 2.20 | 2.05 | 1.79 | 1.44 | 1.69 | 1.04 | 1.52 | 1.96 | 1.19 | 0.92 | 1.51 |
| Transformer-Large | 4.05 | 0.58 | 1.52 | 0.64 | 1.53 | 1.55 | 0.48 | 0.45 | 2.02 | 2.12 | 0.39 | 1.59 | 0.53 | 1.54 | 0.33 | 0.79 | 0.54 | 1.08 | 0.85 | 0.64 | 1.40 |
| Transformer-Base | 1.90 | 0.55 | 1.82 | 0.94 | 0.89 | 1.25 | 1.41 | 0.80 | 1.81 | 2.05 | 0.42 | 1.72 | 0.87 | 1.74 | 0.54 | 1.04 | 0.28 | 1.26 | 0.71 | 1.47 | 1.67 |
| Transformer-Small | 4.11 | 2.38 | 0.52 | 2.70 | 0.54 | 1.28 | 1.52 | 1.05 | 1.43 | 1.72 | 0.47 | 1.48 | 0.41 | 1.35 | 0.52 | 1.08 | 1.49 | 1.37 | 0.49 | 1.00 | 1.22 |

Table 73: Quantifying the cross-/mono-lingual summarization gap. Scores (validation set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.96 | 72.27 | 76.72 | 72.46 | 71.11 | 76.55 | 73.84 |
| Gold + Pseudo (Static) | 74.68 | 72.95 | 77.26 | 72.96 | 72.06 | 76.45 | 74.39 |
| Gold + Pseudo (Dropout) | 74.85 | 72.85 | 77.28 | 72.47 | 72.17 | 76.54 | 74.36 |

Table 74: Effect of dropout on teacher models. Scores (validation set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.96 | 72.27 | 76.72 | 72.46 | 71.11 | 76.55 | 73.84 |
| Gold + Pseudo (Cross-Ensemble) | 74.15 | 72.30 | 76.42 | 72.33 | 70.81 | 75.97 | 73.66 |
| Gold + Pseudo (Mono-Ensemble) | 74.85 | 72.87 | 77.28 | 72.46 | 72.22 | 76.55 | 74.37 |

Table 75: Choice of teacher models. Scores (validation set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.96 | 72.27 | 76.72 | 72.46 | 71.11 | 76.55 | 73.84 |
| Gold + Pseudo (High) | 74.67 | 72.88 | 77.05 | 73.12 | 72.14 | 76.13 | 74.33 |
| Gold + Pseudo (Low) | 74.75 | 72.79 | 77.20 | 72.78 | 71.94 | 76.16 | 74.27 |
| Gold + Pseudo (Diverse) | 75.09 | 73.26 | 77.50 | 72.95 | 72.30 | 76.40 | 74.58 |

Table 76: Characteristics of pseudo-references. Scores (validation set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.96 | 72.27 | 76.72 | 72.46 | 71.11 | 76.55 | 73.84 |
| Gold + Pseudo-1 | 74.45 | 72.55 | 76.88 | 72.80 | 71.88 | 76.16 | 74.12 |
| Gold + Pseudo-32 | 74.64 | 72.52 | 77.10 | 72.42 | 72.39 | 76.29 | 74.23 |
| Gold + Pseudo-96 | 74.85 | 72.85 | 77.28 | 72.47 | 72.17 | 76.54 | 74.36 |

Table 77: Varying the size of the pseudo-reference pool. Scores (validation set) are displayed in BERTScore.

[14] https://huggingface.co/google-bert/bert-base-multilingual-cased

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 73.96 | 72.27 | 76.72 | 72.46 | 71.11 | 76.55 | 73.84 |
| Gold (Soft) | 74.72 | 72.88 | 77.25 | 71.94 | 70.81 | 75.50 | 73.85 |
| Gold (One-Hot) + Pseudo (Soft) | 74.85 | 72.87 | 77.28 | 72.46 | 72.22 | 76.55 | 74.37 |

Table 78: Labeling the gold summary with teacher models. Scores (validation set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.96 | 72.27 | 76.72 | 72.46 | 71.11 | 76.55 | 73.84 |
| Pseudo | 74.95 | 72.84 | 77.43 | 72.30 | 71.38 | 76.11 | 74.17 |
| Gold + Pseudo | 74.85 | 72.87 | 77.28 | 72.46 | 72.22 | 76.55 | 74.37 |

Table 79: Training with only pseudo-labels. Scores (validation set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.86 | 72.25 | 76.78 | 72.33 | 71.31 | 76.15 | 73.78 |
| Gold + Pseudo (Static) | 74.68 | 73.08 | 77.20 | 72.68 | 71.44 | 75.93 | 74.17 |
| Gold + Pseudo (Dropout) | 74.76 | 73.22 | 77.34 | 72.38 | 72.00 | 75.99 | 74.28 |

Table 80: Effect of dropout on teacher models. Scores (test set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.86 | 72.25 | 76.78 | 72.33 | 71.31 | 76.15 | 73.78 |
| Gold + Pseudo (Cross-Ensemble) | 73.94 | 72.40 | 76.56 | 71.96 | 70.50 | 75.48 | 73.47 |
| Gold + Pseudo (Mono-Ensemble) | 74.76 | 73.22 | 77.34 | 72.38 | 72.00 | 75.99 | 74.28 |

Table 81: Choice of teacher models. Scores (test set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.86 | 72.25 | 76.78 | 72.33 | 71.31 | 76.15 | 73.78 |
| Gold + Pseudo (High) | 74.54 | 72.97 | 77.22 | 72.49 | 71.62 | 75.91 | 74.12 |
| Gold + Pseudo (Low) | 74.72 | 72.96 | 77.30 | 72.37 | 71.70 | 75.92 | 74.16 |
| Gold + Pseudo (Diverse) | 75.00 | 73.27 | 77.57 | 72.65 | 71.98 | 76.21 | 74.45 |

Table 82: Characteristics of pseudo-references. Scores (test set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.86 | 72.25 | 76.78 | 72.33 | 71.31 | 76.15 | 73.78 |
| Gold + Pseudo-1 | 74.37 | 72.55 | 76.99 | 72.30 | 71.75 | 75.78 | 73.96 |
| Gold + Pseudo-32 | 74.76 | 72.86 | 77.31 | 72.42 | 71.67 | 76.07 | 74.18 |
| Gold + Pseudo-96 | 74.76 | 73.22 | 77.34 | 72.38 | 72.00 | 75.99 | 74.28 |

Table 83: Varying the size of the pseudo-reference pool. Scores (test set) are displayed in BERTScore.

| | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold (One-Hot) | 73.86 | 72.25 | 76.78 | 72.33 | 71.31 | 76.15 | 73.78 |
| Gold (Soft) | 74.74 | 72.88 | 77.31 | 71.44 | 70.15 | 75.19 | 73.62 |
| Gold (One-Hot) + Pseudo (Soft) | 74.76 | 73.22 | 77.34 | 72.38 | 72.00 | 75.99 | 74.28 |

Table 84: Labeling the gold summary with teacher models. Scores (test set) are displayed in BERTScore.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| Gold | 73.86 | 72.25 | 76.78 | 72.33 | 71.31 | 76.15 | 73.78 |
| Pseudo | 74.87 | 73.08 | 77.53 | 71.73 | 70.50 | 75.61 | 73.89 |
| Gold + Pseudo | 74.76 | 73.22 | 77.34 | 72.38 | 72.00 | 75.99 | 74.28 |

Table 85: Training with only pseudo-labels. Scores (test set) are displayed in BERTScore.

|  | Zh2Tr | Tr2Vi | En2Tr | Tr2En | Ko2Zh | Ko2Ja | Ja2Ru | Ja2Ar | Ar2Ko | En2Ko | Ko2En | En2Ja | Ja2En | En2Zh | Zh2En | En2Vi | Vi2En | En2Ar | Ar2En | En2Ru | Ru2En | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trans-Sum | 62.13 | 67.30 | 62.24 | 69.09 | 64.68 | 59.88 | 66.19 | 64.31 | 61.10 | 61.69 | 68.29 | 58.40 | 68.18 | 64.92 | 68.15 | 67.65 | 68.67 | 65.13 | 66.67 | 65.76 | 68.05 | 65.17 |
| Sum-Trans | 70.42 | 72.39 | 72.08 | 74.24 | 74.34 | 71.79 | 72.82 | 73.76 | 73.78 | 74.25 | 75.54 | 72.43 | 74.97 | 74.79 | 75.95 | 76.96 | 75.72 | 75.06 | 75.07 | 73.99 | 75.40 | 74.08 |
| Gold | 71.31 | 76.15 | 72.33 | 77.02 | 74.61 | 72.25 | 73.73 | 74.27 | 73.86 | 74.57 | 76.78 | 72.78 | 76.46 | 75.40 | 76.26 | 77.39 | 76.05 | 75.36 | 76.12 | 74.55 | 76.21 | 74.93 |
| Gold + OTSum | 70.90 | 75.47 | 72.42 | 76.71 | 75.68 | 72.80 | 74.44 | 74.37 | 74.11 | 74.73 | 75.77 | 73.39 | 75.19 | 75.83 | 76.06 | 77.28 | 77.12 | 74.72 | 75.48 | 74.48 | 75.81 | 74.89 |
| Gold + Many2Many | 71.64 | 77.05 | 72.07 | 78.16 | 75.78 | 72.81 | 74.54 | 74.81 | 74.37 | 74.90 | 77.26 | 72.89 | 77.05 | 75.74 | 76.81 | 77.89 | 77.20 | 75.26 | 76.62 | 74.54 | 76.42 | 75.42 |
| Gold + Pseudo | 72.00 | 75.99 | 72.38 | 77.86 | 75.43 | 73.22 | 74.63 | 75.07 | 74.76 | 75.27 | 77.34 | 73.42 | 77.06 | 75.57 | 76.94 | 77.49 | 77.09 | 75.82 | 76.95 | 74.80 | 76.89 | 75.52 |

Table 86: Fully Supervised Training with mBART-50. Scores (test set) are displayed in BERTScore.

|  | Ar2Ko | Ko2Ja | Avg. |
|---|---|---|---|
| [Transformer] Gold | 71.87 | 69.80 | 70.84 |
| [Transformer] Gold + Pseudo | 72.27 | 70.35 | 71.31 |
| [PISCES] Gold | 74.68 | 73.21 | 73.94 |
| [PISCES] Gold + Pseudo | 75.14 | 73.35 | 74.25 |

Table 87: Fully Supervised Training with other architectures. Scores (test set) are displayed in BERTScore.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [50] Gold | 69.78 | 68.18 | 72.26 | 66.73 | 67.18 | 71.79 | 69.32 |
| [50] Gold + Pseudo | 70.94 | 69.64 | 72.99 | 67.15 | 66.81 | 72.82 | 70.06 |
| [100] Gold | 70.32 | 68.90 | 72.85 | 67.88 | 67.14 | 72.83 | 69.99 |
| [100] Gold + Pseudo | 71.26 | 70.16 | 73.48 | 68.18 | 67.42 | 73.08 | 70.60 |
| [300] Gold | 71.25 | 70.06 | 73.33 | 68.23 | 67.98 | 72.81 | 70.61 |
| [300] Gold + Pseudo | 72.35 | 70.85 | 74.02 | 69.32 | 68.92 | 74.56 | 71.67 |

Table 88: Few-Shot Learning (50/100/300-shot). Scores (test set) are displayed in BERTScore.

|  | Ar2Ko | Ko2Ja | Ko2En | En2Tr | Zh2Tr | Tr2Vi | Avg. |
|---|---|---|---|---|---|---|---|
| [LoRA] Gold | 73.35 | 71.93 | 75.94 | 70.62 | 68.56 | 75.19 | 72.60 |
| [LoRA] Gold + Pseudo | 74.02 | 72.31 | 76.39 | 70.51 | 69.29 | 75.69 | 73.03 |

Table 89: Low-Rank Adaptation. Scores (test set) are displayed in BERTScore.

### D.6 Ensemble Perplexity

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| En | 5.50 | 5.37 | 5.36 | 5.35 | 5.36 | 5.37 |
| Vi | 4.87 | 4.71 | 4.65 | 4.63 | 4.62 | 4.62 |
| Ja | 5.23 | 5.07 | 5.02 | 4.99 | 4.98 | 4.98 |
| Zh | 7.75 | 7.44 | 7.34 | 7.29 | 7.27 | 7.25 |
| Ar | 6.44 | 6.24 | 6.16 | 6.13 | 6.11 | 6.10 |
| Ko | 5.16 | 4.97 | 4.92 | 4.91 | 4.90 | 4.90 |
| Ru | 4.76 | 4.64 | 4.61 | 4.60 | 4.59 | 4.61 |
| Tr | 5.89 | 5.36 | 5.29 | 5.25 | 5.25 | 5.25 |

Table 90: Perplexity scores (lower is better) of ensemble sets on the validation split. Here each row denotes the perplexity score of the best ensemble set (for the given size) of each language.

## E Annotation Instruction

During annotation of a sample, each participant was presented with the source document, gold summary, and the three model-generated summaries which were shuffled beforehand to avoid positional bias.

For *informativeness*, participants were asked to rank model summaries based on whether they captured the most important information (as presented in the gold summary), to which extent was that information covered; and if there was additional information (i.e. not in the gold summary), then based on the theme of the document and content of the gold summary, whether that additional information was also informative or simply redundant, and if any information appeared relevant but contradicted the source document/gold summary (i.e. hallucination), it was treated as harmful and not informative.

For *fluency*, participants were asked to rank model summaries based on how well-formed they were: Was there any grammar, lexical or typographical error ? Were there foreign words mixed in (except for normal keywords that were also in the source document/gold summary) ? Was the summary well-formatted (e.g. no weird next line, random placement of punctuation) ? Was there any weird point (e.g. unnatural but still understandable, or completely absurd sentence) ? Was the summary well-presented (e.g. logically connected and easy-to-follow, which also relate to coherence) ?

Ultimately, participants had to take these factors into consideration and produced the final rankings based on their own estimations. In scenarios where they did not perceive noticeable quality difference in the specified category (*informativeness* or *fluency*) between two (or all three) model summaries, they were allowed to place them in the same ranking at that category. In such scenarios, model summaries of the same ranking would receive similar scores. For example, if there are two model summaries placing second, and one places first, then those placing second would each gain a score of 2 and the one placing first would gain a score of 3. If all model summaries are deemed equal, they each receive a score of 3.

## F Summary Outputs

As case study, we show a random sample from the English-to-Vietnamese (EN2VI) test split accompanied with four model summaries in Table 91. English translations for the Vietnamese summaries are provided underneath in *italics*. We highlight the misleading information in red and the parts aligning with the gold summary in blue. Here we can see that most models produce misinformation. The GOLD summary suggests one simply drink *liquid* (instead of *water*) and generates the word tinh dầu nha chu which in itself is contradicting because tinh dầu refers to *oil* whereas nha chu refers to *periodontitis* which is of a different type. The GOLD+OTSUM summary mistakes dầu đinh hương (*clove oil*) with dầu gội đầu (*shampoo*). The GOLD+MANY2MANY summary additionally includes the phrase không kê toa (*over-the-counter*) which is also misleading because readers easily get the impression that this is the most (or only) suggested way (which is not true as can be inferred from the source document). In addition, it also mistakes dầu đinh hương (*clove oil*) with dầu cây phỉ (*Witch Hazel oil*). Meanwhile, the summary generated by GOLD+PSEUDO aligns well with the gold summary and does not contain misinformation.

| |
|---|
| **Source Document**<br>[.....] Your doctor may recommend a prescription-strength pain medication, or you may wish to stick with over-the-counter medications like aspirin or acetaminophen. Do NOT give aspirin to children or adolescents. Use of aspirin in children or teenagers may cause complications with the liver and brain. [.....] Don't exceed the dosage with ibuprofen either because this can lead to severe stomach or intestinal bleeding. Use cold packs only for the first 48 hours. Fill a sandwich bag with ice cubes, or wrap ice cubes in a clean towel. In a pinch, you can also use a bag of frozen vegetables wrapped in a paper towel. Apply to the affected side of the face. Remove the bag if it starts to feel like it is burning your skin or you may damage your skin. Keep the ice pack on for 20 minutes, then off for 20 minutes. After two days you should switch to using a warm compress, as a cold compress will no longer reduce swelling or inflammation after the first 48 hours. Drinking clear liquids, especially water at room temperature, is crucial following any surgical procedure. Avoid alcohol after any surgery. Water at room temperature is the best beverage to stay hydrated. If you like, you may wish to alternate water with a sugar-free sports drink. This will remove debris and help sooth inflammation. Add approximately half a teaspoon of salt to one cup of lukewarm water. Stir the salt water thoroughly, so that it is mostly dissolved. Very gently swish the salt water around in your mouth, focusing on the affected side of your mouth without creating negative pressure which can dislodge the clot. Repeat after each meal and before bed, and any times in between when you believe a saltwater rinse might be helpful. The physical act of smoking a cigarette may cause a blood clot to come dislodged, and using chewing tobacco or passing smoke over the socket may further irritate the wound and prolong the pain and inflammation. Try using a nicotine patch if you don't believe you can't quit smoking during the time it takes for your mouth to recover. Talk to your doctor about smoking alternatives. Some patients have found that using a few drops of clove oil to treat the site of an open dental wound may alleviate some of the pain. This should not be done as an alternative to seeking professional medical advice and assistance. Rather, it is meant to provide temporary pain relief in the event that immediate medical assistance is not available. Apply one or two drops of clove oil to a clean cotton swab. Dab the swab against your gum where the tooth was extracted. [.....] |
| **Reference Source Summary**<br>Take pain relievers. Use ice or a cold compress on the side of your face experiencing pain. Stay hydrated. Use a salt water rinse. Avoid using tobacco products. Try clove oil. |
| **Reference Target Summary**<br>Uống thuốc giảm đau. Chườm đá hoặc túi chườm lạnh lên bên mặt bị đau. Cung cấp nước cho cơ thể. Súc miệng nước muối. Tránh sử dụng các sản phẩm thuốc lá. Thử dùng dầu đinh hương.<br>*Take pain relievers. Use ice or a cold compress on the side of your face experiencing pain. Provide water for your body. Gargle with salt water. Avoid using tobacco products. Try clove oil.* |
| **Output - Gold**<br>Uống thuốc giảm đau. Chườm đá viên. Uống nhiều chất lỏng. Súc miệng nước muối. Không hút thuốc. Dùng tinh dầu nha chu.<br>*Take pain relievers. Apply ice cubes. Drink plenty of liquid. Gargle with salt water. Do not smoke. Use periodontal essential oil.* |
| **Output - Gold+OTSum**<br>Dùng thuốc giảm đau. Dùng túi chườm lạnh. Uống nhiều nước. Súc miệng bằng nước muối. Không hút thuốc. Dùng dầu gội đầu.<br>*Take pain relievers. Use a cold compress. Drink plenty of water. Gargle with salt water. Do not smoke. Use shampoo.* |
| **Output - Gold+Many2Many**<br>Uống thuốc giảm đau không kê toa. Chườm đá. Uống nhiều nước. Súc miệng nước muối. Tránh hút thuốc. Dùng dầu cây phỉ.<br>*Take over-the-counter pain relievers. Apply ice. Drink plenty of water. Gargle with salt water. Avoid smoking. Use Witch Hazel oil.* |
| **Output - Gold+Pseudo**<br>Uống thuốc giảm đau. Chườm đá viên. Uống nhiều nước. Súc miệng bằng nước muối. Tránh hút thuốc. Dùng dầu đinh hương.<br>*Take pain relievers. Apply ice cubes. Drink plenty of water. Gargle with salt water. Avoid smoking. Use clove oil.* |

Table 91: A random sample from the English-to-Vietnamese (En2Vi) test set