Multilingual Sentence-Level Semantic Search using Meta-Distillation Learning

Anonymous ACL submission

Abstract

Multilingual semantic search is the task of retrieving relevant contents to a query expressed in different language combinations. It is less explored and more challenging than its monolingual or bilingual counterparts, due to the need to circumvent "language bias". Overcoming language bias requires a stronger alignment approach to pull the contents to be retrieved close to the representation of their corresponding queries no matter their language combinations. Traditionally, this is achieved through more supervision in the form of multilingual parallel resources which is expensive to obtain. In this work, we propose a novel alignment approach: MAML-Align,[1] specifically for low-resource multilingual semantic search. Our approach leverages meta-distillation learning on top of MAML, an optimization-based Model-Agnostic Meta-Learner. MAML-Align distills knowledge from a **T**eacher meta-transfer model T-MAML, specialized in transferring from monolingual to bilingual semantic search, to a **S**tudent model S-MAML, which transfers from bilingual to multilingual semantic search. To the best of our knowledge, we are the first to extend meta-distillation to a multilingual search application. Our low-resource evaluation shows that on top of a strong baseline based on sentence transformers, our meta-distillation approach significantly outperforms naive fine-tuning and vanilla MAML.

1 Introduction

The web offers a wealth of information in multiple languages presenting a challenge for reliable, efficient, and accurate information retrieval. Users across the globe may express the need to retrieve relevant content in a language different from the language of the query or in multiple languages simultaneously. These observations bolster the strong demand for multilingual semantic search.

[1]We will release our code in the camera-ready version.

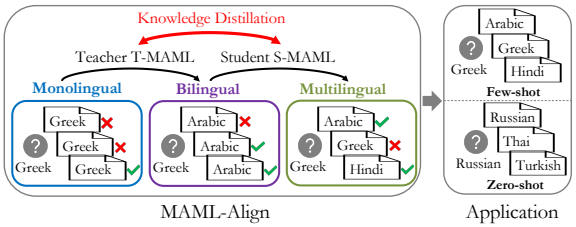


Figure 1: A high-level diagram of our meta-distillation **MAML-Align** framework for multilingual semantic search and some of its application scenarios. This differs from standard cross-lingual transfer setups where the focus is on transferring between individual languages. Given the nature of the downstream task where multiple language combinations could be used in the query and content to be retrieved, we study the transfer here between different variants of the task. As illustrated above, we focus on the three most to least resourced variants where the queries and contents are either from the same language (monolingual), two different languages (bilingual), or multiple languages (multilingual). We leverage knowledge distillation to align between the teacher **T-MAML** (Finn et al., 2017), specialized in transferring from monolingual to bilingual, and the student **S-MAML** specialized in transferring from bilingual to multilingual semantic search. We show the merit of gradually transferring between those variants through few-shot and zero-shot applications involving different language arrangements in the training and evaluation.

Compared to bilingual semantic search, often portrayed as cross-lingual information retrieval (Savoy and Braschler, 2019; Grefenstette, 1998), multilingual semantic search, which involves retrieving answers in multiple languages is under-explored and more challenging. One of the main challenges of multilingual semantic search is the need to circumvent "language bias". Language bias is the tendency of a model to prefer one language over another making it prone to retrieve answers from the preferred language more regardless of how relevant they truly are. For example, a weakly aligned model which clusters relevant content and queries more by language while poorly encapsulating their meaning could pick answers that match the lan-

guage of the query even if they are incorrect. Circumventing language bias requires stronger alignment to factor out language so that the most semantically relevant pairs across languages stand out as closest in the embedding space (Roy et al., 2020).

The majority of approaches used to improve the alignment between languages require parallel resources across languages (Cao et al., 2020; Zhao et al., 2021) which are expensive to obtain especially for multilingual tasks and biased towards high-resource language pairs/combinations. Pre-trained unsupervised multilingual encoders such as M-BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have been employed as off-the-shelf zero-shot tools for cross-lingual and multilingual downstream applications. However, these models still fail to significantly outperform traditional static cross-lingual embeddings (Glavaš et al., 2019) on multilingual semantic search (Litschko et al., 2021). Simply fine-tuning M-BERT and XLM-R on English data is not sufficient to produce an embedding space that exhibits strong alignment (Roy et al., 2020). In fact, fine-tuning such models to largely available monolingual data makes them prone to "monolingual overfitting" as they are shown to transfer reasonably well to other monolingual semantic search settings but not necessarily to bilingual and multilingual settings (Litschko et al., 2022).

Knowledge distillation and contrastive-distillation learning approaches are used to produce better-aligned multilingual sentence representations with reduced need for parallel corpora (Reimers and Gurevych, 2020; Tan et al., 2023). However, they still rely on some supervision in the form of monolingual corpora and back-translation. Cross-lingual meta-transfer learning (Nooralahzadeh et al., 2020; M'hamdi et al., 2021) leveraging MAML (Finn et al., 2017) has been shown to reduce overfitting to high-resource monolingual setups and improve the generalization to new languages with little to no training. However, meta-learning can also be prone to overfitting when multiple source domains or task variants are trained on as part of one single model which undermines its transferring capabilities (Zhong et al., 2022).

To obtain a stronger alignment while preventing monolingual overfitting and with decreased reliance on parallel resources, we propose a low-resource adaptation of meta-distillation learning to multilingual semantic search. We pursue a meta-learning direction based on MAML to allow us to effectively leverage high-resourced monolingual and bilingual variants of semantic search to effectively transfer to multilingual semantic search. To improve the meta-transferring capabilities of MAML, we explore the combination of meta-learning and knowledge distillation (Zhou et al., 2022; Liu et al., 2022; Zhang et al., 2020) and propose a new algorithm for gradually adapting them to the task of multilingual semantic search MAML-Align (Figure 1). We perform MAML-Align in two stages 1) from monolingual to bilingual and 2) from bilingual to multilingual to create a more gradual feedback loop, which makes it easier to generalize to the multilingual case. We conduct experiments on two different semantic search benchmarks: LAReQA (Roy et al., 2020), a span-based question-answering task reformulated as a retrieval task, and $\text{STSB}_{\text{Multi}}$ (Cer et al., 2017), a semantic similarity task. Our experiments show that our multilingual meta-distillation approach beats vanilla MAML and achieves statistically significant gains of 0.6% and 10.6% on LAReQA and 1.2% and 2.5% on $\text{STSB}_{\text{Multi}}$ over an off-the-shelf zero-shot baseline based on sentence transformers (Reimers and Gurevych, 2019) and naive fine-tuning, respectively. We also show consistent gains for both benchmarks on different languages even those kept for zero-shot evaluation. Our approach is model-agnostic and is extensible to other challenging multilingual and cross-lingual downstream tasks requiring strong alignment.

Our **main contributions** are: (1) We are the first to propose a meta-learning approach for multilingual semantic search (§3) and to curate meta-tasks for that effect (§4.2). (2) We are the first to propose a meta-distillation approach to transfer semantic search ability between monolingual, bilingual, and multilingual data (§3.3). (3) We systematically compare between several few-shot transfer learning methods and show the gains of our multilingual meta-distillation approach (§5.1). (4) We also conduct ablation studies involving different language arrangements and sampling approaches (§5.2).

2 Multilingual Semantic Search

In this section, we define sentence-level semantic search and its different categories (§2.1), language variants (§2.2), and supervision degrees (§2.3).

2.1 Task Formulation

Our base task is sentence-level semantic search. Given a sentence query q from a pool of queries $\mathcal{Q}$,

the goal is to find relevant content r from a pool of candidate contents $\mathscr{R}$. The queries are sentences and retrieved contents are either sentences or small passages of a few sentences.

In terms of the format of the queries and contents, there are two main categories of semantic search: (1) **Symmetric Semantic Search.** Each query q and its corresponding relevant content r have similar length and format. (2) **Asymmetric Semantic Search.** q and r are not of the same format. For example, q and r can be a question and a passage answering that, respectively.

2.2 Task Language Variants

In the context of languages, we distinguish between three variants of semantic search at evaluation time (also shown in Figure 1): (1) **Monolingual Semantic Search (mono).** The pools of queries and candidate contents $\mathscr{Q}$ and $\mathscr{R}$ are from the same known and fixed language $\boldsymbol{\ell}_{\mathscr{Q}} = \boldsymbol{\ell}_{\mathscr{R}} \in \mathscr{L}$. (2) **Bilingual Semantic Search (bi).** The pools of queries and candidate contents are sampled from two different languages $\{\boldsymbol{\ell}_{\mathscr{Q}}, \boldsymbol{\ell}_{\mathscr{R}}\} \in \mathscr{L}^2$, such that $\boldsymbol{\ell}_{\mathscr{Q}} \neq \boldsymbol{\ell}_{\mathscr{R}}$. (3) **Multilingual Semantic Search (multi).** This is the problem of retrieving relevant contents from a pool of candidates from a subset of multiple languages $\mathscr{L}_{\mathscr{R}} \subseteq \mathscr{L}$ to a query expressed in a subset of multiple languages $\mathscr{L}_{\mathscr{Q}} \subseteq \mathscr{L}$. Unlike other variants (monolingual and bilingual), multilingual semantic search doesn't restrict which languages can be used in the queries or the candidate contents.

2.3 Supervision Degrees

In the absence of enough training data for the task, we distinguish between three degrees of supervision of semantic search:

- **Zero-Shot Learning.** This resembles ad-hoc semantic search in that it doesn't involve any fine-tuning specific to the task of semantic search. Rather, off-the-shelf pre-trained language models are used directly to find relevant content to a specific query. This still uses some supervision in the form of parallel sentences used to pre-train those off-the-shelf models. In the context of multilingual semantic search, we include in the zero-shot learning case any evaluation on languages not seen during fine-tuning.
- **Few-Shot Learning.** Few-shot learning is used in the form of a small fine-tuning dataset. In the context of multilingual semantic search, few-shot learning on a particular language implies that that language is seen during fine-tuning or meta-learning either to represent the query or the contents to be retrieved.

3 Multilingual Meta-Distillation Learning

In this section, we start by giving some background on meta-learning (§3.1) and the original MAML algorithm (§3.2), then we present our optimization-based meta-distillation learning algorithm MAML-Align (§3.3) and how it differs from MAML.

3.1 Meta-Learning Background

Meta-learning is a techniques used for fast adaptation to new domains, tasks, and languages. This is done by repeatedly simulating the learning process on the target tasks using many high-resource ones (Gu et al., 2018). The main distinction between meta-learning and conventional machine learning is that while the latter focuses on one data instance at a time, the former optimizes over a distribution of many sub-tasks, referred to as 'meta-tasks', sampled to simulate a low-resource scenario. Each meta-task is defined as a tuple $T_i = (S_i, Q_i)$, where S_i and Q_i denote support and query sets, respectively. Each S_i and Q_i are labeled samples from the downstream task data. In bi-level optimization approaches (which we focus on in this paper), the meta-learner trains on the support set in the inner loop to produce a learner that will make predictions on the query set, and then use that to update the meta-parameters in the outer loop. Therefore, the inner loop is specialized in learning task-specific optimizations over the support sets; the outer loop, on the other hand, learns the generalization over the query sets in a leader-follower manner (Hospedales et al., 2020).

3.2 Original MAML Algorithm

Our first variant is a direct adaptation of MAML to multilingual semantic search. We use the procedure outlined in Algorithm 1. We start by sampling a batch of meta-tasks from a meta-dataset distribution $\mathcal{D}_{X \rightarrow X\prime}$, which simulates the transfer from X to $X\prime$. X and $X\prime$ are different task language variants of semantic search (§2.2) from which the support and query sets are sampled, respectively. We start by initializing our meta-learner parameters θ with the pre-trained base model parameters θ_B. For each batch of meta tasks, we perform an inner loop (Algorithm 2): we go over each meta-task $T_j = (S_j, Q_j)$ in $\mathcal{T}$ where we update θ_j using S_j^X. After n steps of this update, we pre-compute the loss of θ_j on $Q_j^{X\prime}$ and save it for later. At the end of all meta-tasks in the batch, we perform

Algorithm 1 MAML: Transfer Learning from X to $X\prime$ (X$\rightarrow X\prime$)

Require: Meta-task set distribution $\mathcal{D}_{\text{X}\rightarrow X\prime}$ simulating transfer from X to $X\prime$ task language variants, pre-trained downstream base model B with parameters θ_B, and meta-learner M with parameters (θ, α, β, n).

Initialize $\theta \leftarrow \theta_B$
while not done **do**
  Sample a batch of meta-tasks $\mathcal{T} = \{T_1, \ldots, T_b\} \sim \mathcal{D}_{\text{X}\rightarrow X\prime}$
  $\sum \mathcal{L}_{\mathcal{T}}^{S^X}, \sum \mathcal{L}_{\mathcal{T}}^{Q^{X\prime}}$ = INNER_LOOP($\mathcal{T}$, θ, α, n)
  Outer Loop: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum \mathcal{L}_{\mathcal{T}}^{Q^{X\prime}}$
end while

Algorithm 2 INNER_LOOP

function INNER_LOOP($\mathcal{T}$, θ, α, n)
  for each $T_j = (S_j^X, Q_j^{X\prime})$ in $\mathcal{T}$ **do**
    Initialize $\theta_j \leftarrow \theta$
    for $t = 1 \ldots n$ **do**
      Evaluate $\partial B_{\theta_j} / \partial \theta_j = \nabla_{\theta_j} \mathcal{L}_{T_j}^{S_j^X}(B_{\theta_j})$
      Update $\theta_j = \theta_j - \alpha \partial B_{\theta_j} / \partial \theta_j$
    end for
    Evaluate query loss $\mathcal{L}_{T_j}^{Q_j^{X\prime}}(B_{\theta_j})$ and save it for outer loop
  end for
  $\sum \mathcal{L}_{\mathcal{T}}^{S^X} \leftarrow \sum_{j=1}^{b} \mathcal{L}_{T_j}^{S_j^X}(B_{\theta_j})$
  $\sum \mathcal{L}_{\mathcal{T}}^{Q^{X\prime}} \leftarrow \sum_{j=1}^{b} \mathcal{L}_{T_j}^{Q_j^{X\prime}}(B_{\theta_j})$
  return $\sum \mathcal{L}_{\mathcal{T}}^{S^X}, \sum \mathcal{L}_{\mathcal{T}}^{Q^{X\prime}}$
end function

one outer loop by summing over all pre-computed gradients and updating θ. Following X-METRADA (M'hamdi et al., 2021), we perform this algorithm in two stages: meta-train and meta-valid where meta-valid is a replication of meta-train with the main difference being the task language variant arrangements used to sample the meta-tasks.

3.3 MAML-Align Algorithm

The idea behind this extension is to use knowledge distillation to distill T-MAML to S-MAML where T-MAML and S-MAML are replicates of MAML and T-MAML is more high-resource than S-MAML. Inspired by M'hamdi et al. (2021) work which shows that multiple phases of bi-level optimization encourages faster adaptation to low-resource languages, we also adopt a gradual approach to meta-transfer across different task language variants with the help with knowledge distillation. Given meta-tasks from $\mathcal{D}_{\text{X}\rightarrow\text{Y}}$ and $\mathcal{D}_{\text{Y}\rightarrow\text{Z}}$, the goal is to use that shared task language variant of transfer Y to align different modes of transfer of semantic search. We start by executing the two inner loops of the two MAMLs (with more inner steps for T-MAML than S-MAML), where the support sets

Algorithm 3 MAML-Align: Knowledge distillation to align two different MAMLs (X$\rightarrow$Y$\rightarrow$Z)

Require: Meta-task set distributions $\mathcal{D}_{\text{X}\rightarrow\text{Y}}$ and $\mathcal{D}_{\text{Y}\rightarrow\text{Z}}$ sharing the same Y, pre-trained downstream base model B with parameters θ_B, and meta-learners $M_{\text{X}\rightarrow\text{Y}}$ with parameters (θ, α, β, n) and $M_{\text{Y}\rightarrow\text{Z}}$ with parameters ($\theta\prime$, α, β, $n\prime$), where $n\prime < n$.

Initialize $\theta \leftarrow \theta_B$
Initialize $\theta\prime \leftarrow \theta_B$
while not done **do**
  Sample batch of tasks $\mathcal{T} \sim \mathcal{D}_{\text{X}\rightarrow\text{Y}}$
  Sample batch of tasks $\mathcal{T}\prime \sim \mathcal{D}_{\text{Y}\rightarrow\text{Z}}$
  $\sum \mathcal{L}_{\mathcal{T}}^{S^X}, \sum \mathcal{L}_{\mathcal{T}}^{Q^Y}$ = INNER_LOOP($\mathcal{T}_{\text{X}\rightarrow\text{Y}}$, θ, α, n)
  $\sum \mathcal{L}_{\mathcal{T}\prime}^{S^Y}, \sum \mathcal{L}_{\mathcal{T}\prime}^{Q^Z}$ = INNER_LOOP ($\mathcal{T}_{\text{Y}\rightarrow\text{Z}}$, $\theta\prime$, α, $n\prime$)
  $\mathcal{L}_{task} = (\sum \mathcal{L}_{\mathcal{T}}^{Q^Y} + \mathcal{L}_{\mathcal{T}\prime}^{Q^Z})/2$
  $\mathcal{L}_{kd} = (\sum \mathcal{L}_{\mathcal{T}}^{Q^Y} - \sum \mathcal{L}_{\mathcal{T}\prime}^{S^Y})^2$
  Update $\theta \leftarrow \theta - \beta \nabla_\theta (\mathcal{L}_{task} + \lambda \mathcal{L}_{kd})$
end while

are sampled from X and Y, respectively. Then, we compute, in the optimization process of the outer loop, the weighted combination of $\mathcal{L}_{task}$, the average over the task-specific losses on the query sets sampled from Y and Z, and $\mathcal{L}_{kd}$, the mean-squared error on Y. Figure 2 illustrates a conceptual comparison between MAML and MAML-Align.

4 Experimental Setup

In this section, we describe the downstream datasets and models used (§4.1), their formulation as meta-tasks (§4.2), and the different baselines and model variants used in the evaluation (§4.3).

4.1 Downstream Benchmarks

We evaluate our proposed approaches over the following multilingual and bilingual sentence-level semantic search datasets for which we describe the downstream models used:[2]

- **Asymmetric Semantic Search.** We use LAReQA (Roy et al., 2020), focusing on XQuAD-R, which is a retrieval-based task reformulated from the span-based question answering XQuAD (Artetxe et al., 2020). This dataset covers 11 languages. In this work, we only use seven languages. Arabic, German, Greek, and Hindi are used for few-shot learning. Russian, Thai, and Turkish are kept for zero-shot evaluation. There are less than 1200 questions and 1300 candidates for each language.[3] We design a Transformer-based triplet-encoder model (modified from the original dual encoder in Roy et al.

[2] More details on the base model architectures can be found in Appendix B. More experimental details on the datasets statistics and hyperparameters used in Appendix C.

[3] We download the data from https://github.com/google-research-datasets/lareqa.

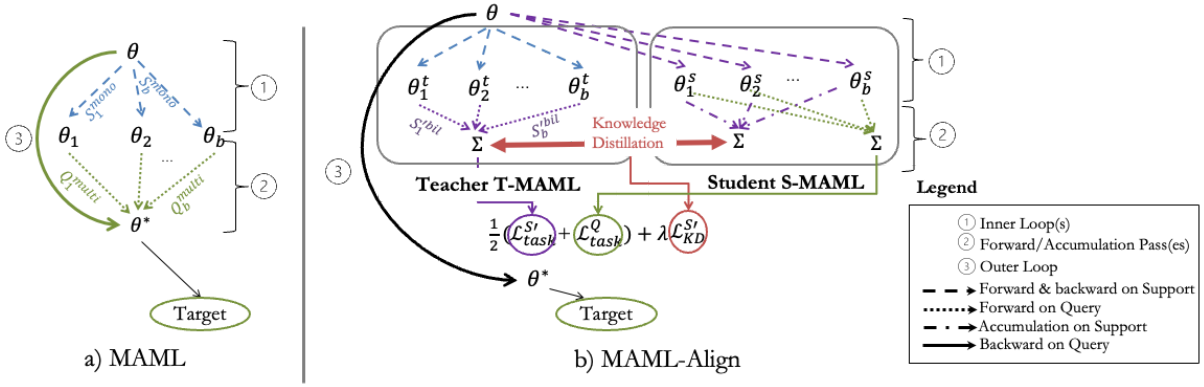


Figure 2: A conceptual comparison between **MAML-Align** and the original meta-learning baseline **MAML**. A single iteration of MAML involves one inner loop optimizing over a batch of support sets from a source language variant of the task followed up by an outer loop optimizing over the batch query sets curated from the target task variant. In MAML-Align, on the other hand, we curate two support sets and one query set, where the second support set is used as both a query and support set in T-MAML and S-MAML, respectively. We perform two inner loops. Then, in the outer loop, we optimize jointly over the distillation and task-specific losses of the query sets.

(2020)) with three towers encoding 1) the question, 2) its answer and its context, and 3) the negative candidates and their contexts. Then, we use triplet loss (Schroff et al., 2015) to minimize the distance between towers 1 and 2 on one hand and maximize the distance between towers 1 and 3 on the other hand.

- **Symmetric Semantic Search.** As there is no multilingual parallel benchmark for symmetric search, we focus, in our few-shot learning experiments, on a small-scale bilingual benchmark. We use STSB$_{\text{Multi}}$ from SemEval-2017 Task 1 (Cer et al., 2017).[4] This is a semantic similarity benchmark, which consists of a collection of sentence pairs drawn mostly from news headlines. It covers English-English, Arabic-Arabic, Spanish-Spanish, Arabic-English, Spanish-English, and Turkish-English. There are only 250 sentence pairs for each language pair. Each sentence pair is scored between 1 and 5 to denote the extent of their similarity. We use a Transformer-based dual-encoder, which encodes sentences 1 and 2 in each sentence pair using a shared encoder. We then compute the cosine similarity score between the encodings of sentences 1 and 2.

4.2 Meta-Datasets

Following our formulation of downstream semantic search benchmarks, we independently construct the support set S in each meta-task by sampling a batch of k question/answer/negative candidates triplets and sentence pairs in LAReQA and STSB$_{\text{Multi}}$, respectively. Then, we construct q triplets or sentence pairs in the query set Q by picking for each triplet or sentence pair in S either a similar or random triplet or sentence pair.[5]

4.3 Baselines & Model Variants

Since we are the first, to the best of our knowledge, to explore meta-learning for bilingual or multilingual information retrieval or semantic search, we only compare with respect to our internal variants and design some external non-meta-learning baselines. We are also the first to explore fine-tuning and meta-learning on extremely small-scale data using cross-validation splits on both benchmarks. This makes it hard to compare with existing approaches, therefore we rely more on our own internal baselines.

Baselines. We design the following baselines:

- *Zero-Shot*: This is our initial zero-shot approach based on an off-the-shelf pre-trained language model. Based on our preliminary performance evaluation of different existing and state-of-the-art off-the-shelf language models in Table 5, we use the best model on our 5-fold cross-validation test splits, which is sentence-BERT (S-BERT) as our zero-shot model.[6]
- *Fine-tune*: On top of our off-the-shelf zero-shot baseline S-BERT, we fine-tune jointly and directly on the support and query sets of each meta-task in both meta-train and meta-valid. This few-shot baseline makes for a fair comparison with the meta-learning approaches.

[4]Downloaded from https://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools.

[5]Details of transfer modes and their support and query set language arrangements are in Appendix C.2.

[6]paraphrase-multilingual-mpnet-base-v2 in https://huggingface.co/sentence-transformers.

Internal Variants. We design the following meta-learning variants:

- *MAML*: On top of S-BERT, we apply MAML (Algorithm 1). At each episode, we conduct a meta-train followed by a meta-valid stage.
- *MAML-Align*: On top of S-BERT, we apply MAML-Align (following Algorithm 3).

External Evaluation. To assess the impact of using machine translation models with or without meta-learning and the impact of machine translation from higher-resourced data, we explore Translate-Train (T-Train), where we translate English data in SQUAD_{EN}[7] and STSB_{EN}[8] to the evaluation languages. We then either use translated data in all languages or in each language separately as a data augmentation technique.

5 Results & Analysis

This section presents the results obtained using different meta-learning model variants compared to the baselines. Given the extremely small-scaled dataset we are working with (Tables 2 and 3), all experiments are evaluated using 5-fold cross-validation and the mean is reported. Following XTREME-R (Ruder et al., 2021) and SemEval-2017 (Cer et al., 2017), scores are reported using mean average precision at 20 (**mAP@20**) and Pearson correlation coefficient percentage (**Pearson's r × 100**) for LAReQA and $\text{STSB}_{\text{Multi}}$, respectively.[9]

5.1 Multilingual Performance Evaluation

Table 1 summarizes the multilingual performances across different baselines and model variants for both semantic search benchmarks. On average, we notice that MAML-Align achieves better results than MAML or S-BERT zero-shot base model and significantly better than Fine-tune. It is worth noting that we report the results for MAML using trans mode, which is trained over a combination of mono→bi and bi→multi in the meta-training and meta-validation stages, respectively. This suggests that MAML-Align helps more in bridging the gap between those transfer modes. We perform a paired two-sample for means t-Test and find that the gains using MAML-Align are statistically significant with p-values of 0.00213 and 0.00248 compared to S-BERT and MAML respectively on LAReQA, rejecting the null hypothesis with 95% confidence.[10] We also observe that fine-tuning baselines are consistently weak compared to different meta-learning model variants, especially for LAReQA. We conjecture that fine-tuning is overfitting to the small amounts of training data, unlike meta-learning approaches which are more robust against that. However, for $\text{STSB}_{\text{Multi}}$, the gap between fine-tuning and meta-learning while still existing and in favor of meta-learning is a bit reduced. We hypothesize that even meta-learning models are suffering from meta-overfitting to some degree in this case.

We notice that T-Train+MAML on top of machine-translated data doesn't necessarily boost the performance on LAReQA or $\text{STSB}_{\text{Multi}}$ on average. This suggest that not all languages used in the machine-translated data provide an equal boost to the performance due to noisy translations for certain languages. While introducing higher quality machine translations could be beneficial in general, there is a compromise to be made in terms of translation API calls overheads and human labor to evaluate the quality of the translations. The purpose

Model	LAReQA	$\text{STSB}_{\text{Multi}}$
Zero-Shot	57.0	<u>81.4</u>
Few-Shot Learning		
Fine-tune	47.0	79.9
MAML(*)	<u>57.2</u>	81.3
MAML-Align(*)	**57.6**	**82.4**
Machine-Translation		
T-Train+Fine-tune	46.1	73.7
T-Train+MAML(*)	57.0	80.9

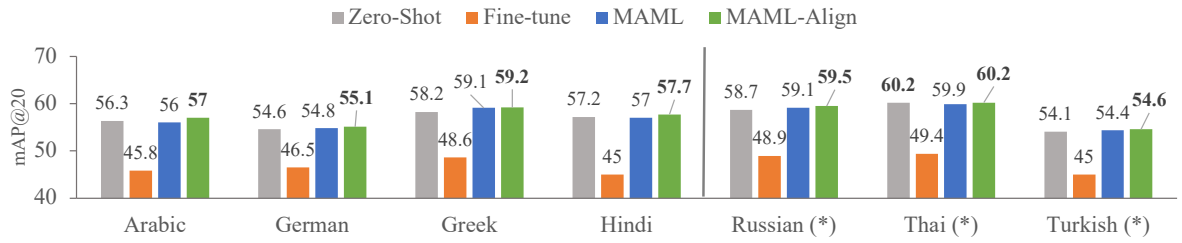
Table 1: This is a comparison of different zero-shot baselines, few-shot learning, and machine translation-enhanced models. Other zero-shot external models (Table 5) show sub-optimal results so we don't include them. For LAReQA and $\text{STSB}_{\text{Multi}}$, we report mAP@20 and Pearson's r × 100, respectively. All results are averaged over 5-fold cross-validation and multiple language choices. Models in (*) are our main contribution. We report the average over many model variants translating from English to one target language at a time for T-Train model variants. Best and second-best results for each benchmark are in **bold** and <u>underlined</u>, respectively.

[7] We use the translate.pseudo-test provided for XQuAD dataset by XTREME (Hu et al., 2020) https://console.cloud.google.com/storage/browser/xtreme_translations.

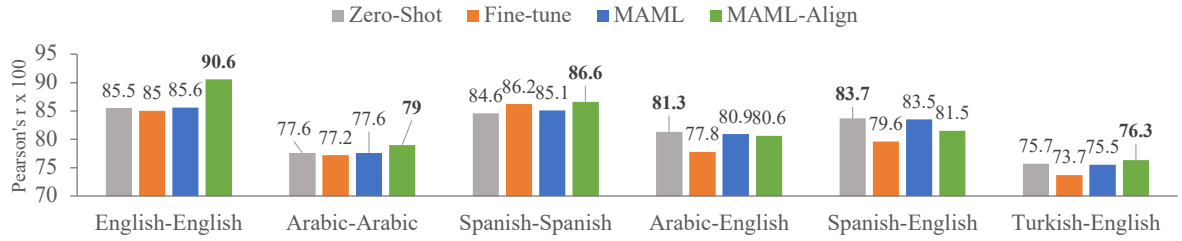
[8] We use the translated dataset from the original English STSB https://github.com/PhilipMay/stsb-multi-mt/.

[9] More fine-grained results for all languages and for both benchmarks can be found in Tables 7 and 8 in Appendix D.

[10] We obtain p-values results 0.0134 and $7.04e-10$ when comparing MAML-Align to S-BERT and MAML using paired t-test on top of bootstrap sampling on the results of each query before taking the mean. The gains using MAML-Align are uniformly consistent for different cross-validation splits.



(a) On LAReQA.



(b) On STSB$_{\text{Multi}}$.

Figure 3: mAP@20 and Pearson's r × 100 5-fold cross-validated multilingual performance evaluation evaluated on LAReQA and STSB$_{\text{Multi}}$ in the first and second subplots, respectively. There are consistent gains in favor of MAML and MAML-Align compared to their fine-tuning and Zero-Shot counterparts for all languages and language-pairs. Languages in (*) are used for zero-shot evaluation, whereas other languages are included either during Meta-train and Meta-valid stages or fine-tuned on. Best results for each language or language pair are highlighted in **Bold**.

of this work is to evaluate in few-shot learning scenarios rather than using data augmentation for that effect. We conjecture that based on our observation in this few-shot learning setup, meta-learning on top of higher-quality machine-translated data could boost the performance even more.

Figure 3 highlights a fine-grained comparison between different model categories on all languages and language pairs for each benchmark. We notice that the gain in favor of meta-learning approaches is consistent across different languages and language pairs. This confirms our findings that while MAML improves a bit over Zero-Shot reducing the impact of overfitting that vanilla Fine-tune suffers from, MAML-Align boosts the gains of meta-learning on all languages and language pairs except for Arabic-English and Spanish-English. The gain applies to zero-shot languages such as Russian and Turkish.

5.2 Ablation Studies

Due to the lack of parallelism in STSB$_{\text{Multi}}$ making a multilingual evaluation on it not possible, we focus hereafter on LAReQA in the remaining analysis and ablation studies. Figure 4 shows the results across different modes of transfer for Fine-tune and MAML. Among all transfer modes, trans, mono→bi, and mono→mono have the best gains, whereas bi→multi and mixt are the weakest forms of transfer. Trans, which uses mono→bi during meta-train and bi→multi during meta-valid, is the best transfer mode for MAML while being one of the weakest for Fine-tune. This not only shows that curating different transfer modes for different meta-learning processes is beneficial but it also suggests that meta-learning is more effective at multi-stage adaptation than fine-tuning on them jointly. Mixt is weaker than trans and this implies that jointly optimizing different forms of transfers of meta-tasks in one stage makes it harder for MAML to learn to generalize. MAML-Align is shown to be better for combining different optimization objectives.

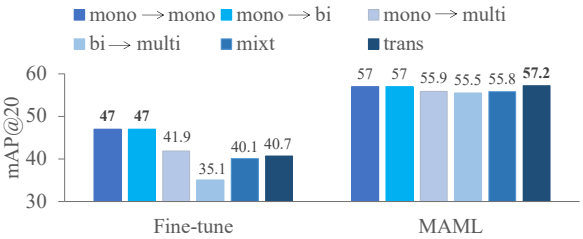


Figure 4: mAP@20 multilingual performance averaged over 5-fold cross-validation splits on LAReQA comparing between different meta-transfer modes for Fine-tune and MAML models. The gap is large between Fine-tune and MAML across all meta-transfer modes and is even larger in favor of MAML when trans mode (uses mono→bi and bi→multi in the meta-training and meta-validation, respectively) is used.

Figure 5 shows a multilingual performance comparison between different sampling modes in meta-tasks constructions. In each meta-task, we either sample the query set that is the most similar to its

corresponding support set (*Similar*) or randomly (*Random*). We hypothesize that the sampling approach plays a role in stabilizing the convergence and generalization of meta-learning. While we were expecting that sampling for each support set a query set that is the most similar to it would help meta-learning converge faster and thus generalize better, it generalized worse on the multilingual performance in this case. On the other hand, random sampling generalizes better to out-of-sample test distributions leading to lower biases between languages in the multilingual evaluation mode.

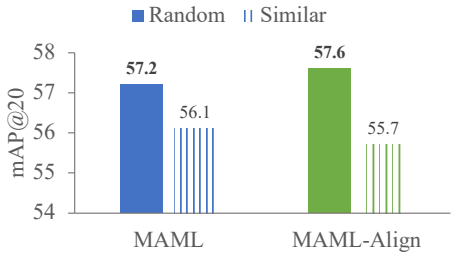


Figure 5: mAP@20 multilingual 5-fold cross-validated performance on LAReQA between different query set sampling modes in meta-tasks for MAML and MAML-Align. We notice that random query sampling has better generalization for both models.

Figure 6 shows the results for different sampling modes of negative examples in the triplet loss. For each support and query set in each meta-task, we either sample random, hard, or semi-hard triplets to test the added value of triplet sampling in few-shot learning. We follow the same approach outlined in Schroff et al. (2015) to sample hard and semi-hard triplets.[11] While we expect training with more hard triplets to help converge the triplet loss in MAML, the multilingual performance using this type of sampling falls short of random sampling. This is due to the fact that more sophisticated ways of triplet loss sampling usually require a more careful hyper-parameter tuning to pick the right amount of triplets. For few-shot learning applications, this usually results in a significant reduction in the number of training examples, which could further hurt the generalization performance. In future work, we plan to investigate hybrid sampling approaches to monitor at which point in meta-learning the training should focus more on hard or easy triplets. This could be done by proposing a regime for making the sampling of meta-tasks dynamic and flexible to also combat meta-over-fitting.

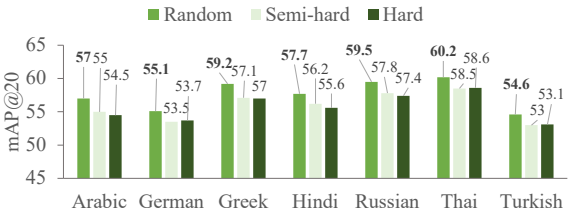


Figure 6: mAP@20 5-fold cross-validated mean multilingual performance over different triplet negative sampling modes on LAReQA tested on different languages using MAML-Align. Random sampling seems best on average for few-shot learning, whereas hard sampling is more stable across cross-validation splits.

6 Related Work

Most approaches to bilingual semantic search rely on machine translation to reduce the problem to monolingual search (Lu et al., 2008; Nguyen et al., 2008; Jones et al., 2008). However, such systems are inefficient due to error propagation and overheads from API calls. Moreover, the number of language combinations in the query and content to be retrieved can get prohibitively large (Savoy and Braschler, 2019). Multilingual models M-BERT and XLM are used for semantic search variants (Yang et al., 2020; Hoogeveen et al., 2015; Lei et al., 2016) but are suboptimal necessitating more parallel resources. Cross-lingual meta-transfer learning applications include Gu et al. (2018); Hsu et al. (2020); Winata et al. (2020); Xiao et al. (2021). Meta-distillation learning has been leveraged either to help the teacher transfer better to the student (Zhou et al., 2022; Liu et al., 2022) or make meta-learning algorithms more portable (Zhang et al., 2020). Xu et al. (2021) follows a gradual multi-stage process which is different from our work in that it uses fine-tuning for domain adaptation to interpolate between in-domain and out-domain data. In contrast, we apply meta-distillation to multilingual semantic search and show that it outperforms gradual fine-tuning.[12]

7 Conclusion

In this work, we adapt multilingual meta-transfer learning combining MAML and knowledge distillation to multilingual semantic search. Our experiments show that our multilingual meta-knowledge distillation approach outperforms both vanilla MAML and fine-tuning approaches on top of a strong sentence transformers model. We evaluate comprehensively on two types of multilingual semantic search and show improvement over the baselines even on unseen languages.

[11] More details about that can also be found in Appendix B.

[12] More detailed related work can be found in Appendix A.

Limitations

In this paper, we have focused on improving multilingual sentence transformers using meta-distillation learning for semantic search. Since our approach tests for strong alignment and is based at its core on a model-agnostic algorithm (MAML), we conjecture that it should be extensible to any multilingual task requiring strong alignment. The community is welcome to further investigate its performance to other benchmarks such as XTREME (Hu et al., 2020) and XTREME-R (Ruder et al., 2021).

All insights and claims from this study are tied to the experimental setup that we describe extensively in the main paper and appendix. We follow a consistent configuration of the hyperparameters for each of the two downstream tasks which we deem to be a fair comparison across all setups and model variants. We don't think that exploring all different combinations of languages in the construction of the query and the content to be retrieved is feasible. So, we leave performing extensive hyperparameters search for different model variants, modes of transfer, and language combinations for future exploration.

We also have memory constraints when it comes to training meta-learning algorithms to deal with ranking and retrieval of sentences from multiple languages at the same time for one query. Our memory constraints make it challenging to explore more sophisticated state-of-the-art Sentence Transformers such as sentence-T5 or GPT Sentence Embeddings SGPT (Ni et al., 2022; Muennighoff, 2022). Applying MAML as an upstream model on top of T5-based downstream model makes it even more computationally infeasible. Our main goal is not to reach state-of-the-art performance for different benchmarks but to showcase the relative advantage of meta-distillation in a few-shot learning setup. Our upstream approach is model-agnostic and can be continuously adapted to novel embedding approaches as they evolve.

References

Sébastien M. R. Arnold, Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias. 2020. learn2learn: A library for meta-learning research.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Vitor R Carvalho, Jonathan L Elsas, William W Cohen, and Jaime G Carbonell. 2008. A meta-learning approach for robust rank learning. In *SIGIR 2008 workshop on learning to rank for information retrieval*, volume 1.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Yi-Chen Chen, Jui-Yang Hsu, Cheng-Kuang Lee, and Hung-yi Lee. 2020. DARTS-ASR: differentiable architecture search for multilingual speech recognition and adaptation. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 1803–1807. ISCA.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Gregory Grefenstette. 1998. Cross language information retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas: Tutorial Descriptions*, Langhorne, PA, USA. Springer.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015, Parramatta, NSW, Australia, December 8-9, 2015*, pages 3:1–3:8. ACM.

Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. 2020. Meta-learning in neural networks: A survey. *CoRR*, abs/2004.05439.

Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee. 2020. Meta learning for end-to-end low-resource speech recognition. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7844–7848. IEEE.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. 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 4411–4421. PMLR.

Gareth Jones, Fabio Fantino, Eamonn Newman, and Ying Zhang. 2008. Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia. In *Proceedings of the 2nd workshop on Cross Lingual Information Access (CLIA) Addressing the Information Need of Multilingual Societies*.

Doron Laadan, Roman Vainshtein, Yarden Curiel, Gilad Katz, and Lior Rokach. 2019. Rankml: a meta learning-based approach for pre-ranking machine learning pipelines. *ArXiv preprint*, abs/1911.00108.

Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis, and Ekaterina Shutova. 2022. Meta-learning for fast cross-lingual adaptation in dependency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8503–8520, Dublin, Ireland. Association for Computational Linguistics.

Hung-yi Lee, Shang-Wen Li, and Thang Vu. 2022. Meta learning for natural language processing: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 666–684, Seattle, United States. Association for Computational Linguistics.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez. 2016. Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1279–1289, San Diego, California. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Chong-En Lin and Kuan-Yu Chen. 2020. A preliminary study on using meta-learning technique for information retrieval. In *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, pages 59–71, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Robert Litschko, Ivan Vulic, Simone Paolo Ponzetto, and Goran Glavas. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 342–358. Springer.

Robert Litschko, Ivan Vulic, Simone Paolo Ponzetto, and Goran Glavas. 2022. On cross-lingual retrieval with multilingual text encoders. *Inf. Retr. J.*, 25(2):149–183.

Jihao Liu, Boxiao Liu, Hongsheng Li, and Yu Liu. 2022. Meta knowledge distillation. *ArXiv preprint*, abs/2202.07940.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Chengye Lu, Yue Xu, and Shlomo Geva. 2008. Web-based query translation for English-Chinese CLIR. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 13, Number 1, March 2008: Special Issue on Cross-Lingual Information Retrieval and Question Answering*, pages 61–90.

Meryem M'hamdi, Doo Soon Kim, Franck Dernoncourt, Trung Bui, Xiang Ren, and Jonathan May. 2021. X-METRA-ADA: Cross-lingual meta-transfer learning adaptation to natural language understanding and question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3617–3632, Online. Association for Computational Linguistics.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search. *ArXiv preprint*, abs/2202.08904.

Pandu Nayak. 2019. Understanding searches better than ever before.

Dong Nguyen, Arnold Overwijk, Claudia Hauff, Dolf Trieschnigg, Djoerd Hiemstra, and Franciska de Jong. 2008. Wikitranslate: Query translation for cross-lingual information retrieval using only wikipedia. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, volume 5706 of *Lecture Notes in Computer Science*, pages 58–65. Springer.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacques Savoy and Martin Braschler. 2019. *Lessons Learnt from Experiments on the Ad Hoc Multilingual Test Collections at CLEF*, pages 177–200. Springer International Publishing, Cham.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1477–1490, Dubrovnik, Croatia. Association for Computational Linguistics.

Ishan Tarunesh, Sushil Khyalia, Vishwajeet Kumar, Ganesh Ramakrishnan, and Preethi Jyothi. 2021.

Meta-learning for effective multi-task and multilingual modelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3600–3612, Online. Association for Computational Linguistics.

Niels van der Heijden, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2021. Multilingual and cross-lingual document classification: A meta-learning approach. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1966–1976, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung. 2020. Meta-transfer learning for code-switched speech recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3770–3776, Online. Association for Computational Linguistics.

Yubei Xiao, Ke Gong, Pan Zhou, Guolin Zheng, Xiaodan Liang, and Liang Lin. 2021. Adversarial meta sampling for multilingual low-resource speech recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14112–14120. AAAI Press.

Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, and Kenton Murray. 2021. Gradual fine-tuning for low-resource domain adaptation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 214–221, Kyiv, Ukraine. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Min Zhang, Donglin Wang, and Sibo Gai. 2020. Knowledge distillation for model-agnostic meta-learning. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1355–1362. IOS Press.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, *SEM 2021, Online, August 5-6, 2021*, pages 229–240. Association for Computational Linguistics.

Tao Zhong, Zhixiang Chi, Li Gu, Yang Wang, Yuanhao Yu, and Jin Tang. 2022. Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts. In *NeurIPS*.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. 2022. BERT learns to teach: Knowledge distillation with meta learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7037–7049, Dublin, Ireland. Association for Computational Linguistics.

Jeffrey Zhu, Mingqin Li, Jason Li, and Cassandra Odoula. 2021. Bing delivers more contextualized search using quantized transformer inference on nvidia gpus in azure.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In *11th Workshop on Building and Using Comparable Corpora - Special Topic: Comparable Corpora for Asian Languages, BUCC@LREC 2018, Miyazaki, Japan, May 8, 2018*. European Language Resources Association.

A More Related Work

Given the scarcity of research on multilingual semantic search using meta-learning and knowledge distillation, we analyze independently previous work in the area of semantic search in general, multilingual semantic search, bilingual meta-transfer learning, and meta-distillation learning before delving into some applications of meta-transfer learning for retrieval ranking and how our work applies meta-transfer and meta-distillation for multilingual semantic search as a whole.

Textual Semantic Search Textual semantic search is the task of retrieving semantically relevant content for a given query. Unlike traditional keyword-matching information retrieval, semantic search seeks to improve search accuracy by understanding the searcher's intent and disambiguating the contextual meaning of the terms in the query (Muennighoff, 2022). Semantic search has broad applications in search engines such as Google (Nayak, 2019), Bing (Zhu et al., 2021), etc. They rely on Transformers (Vaswani et al., 2017) as their dominant architecture going beyond non-semantic models such as BM25 (Robertson and Zaragoza, 2009).

Multilingual Semantic Search Previous work which extends semantic search to different languages is often focused on cross-lingual information retrieval. Progress in cross-lingual information retrieval (CLIR) or semantic search has seen multiple waves (Grefenstette, 1998). Traditionally, when we think of CLIR we automatically think of machine translation (MT) as if they are two faces to the same coin. The only difference is that translation tools are used to render documents readable in the case of MT whereas CLIR focuses on rendering them searchable if at the very core translation technology is what is used for CLIR and MT rather than other paradigms such as transfer learning. Most approaches that fall into this category translate queries into the language of the documents and then perform monolingual search (Lu et al., 2008; Nguyen et al., 2008; Jones et al., 2008). While this is an efficient option, that might not be the most effective approach as queries can be so short and ungrammatical making them hard to translate accurately. So, in this case, translating all documents or sentences to the target languages can be used leading to better accuracy but less efficiency. This translation form is even more inefficient in the case of multilingual semantic search where the number of possible language combinations that can be used in the source and target languages can grow exponentially. Those pipeline approaches suffer from error propagation of the machine translation component into the downstream semantic search, especially for low-resource languages.

More prominent approaches include transfer learning where both query and documents or sentences are encoded into a shared space. The first class of approaches in this category use pre-trained language models where both the query and the documents are encoded into a shared space. The cross-lingual ability of models like M-BERT and XLM has been analyzed for different retrieval-based downstream applications including question-answer retrieval (Yang et al., 2020), bitext mining (Ziemski et al., 2016; Zweigenbaum et al., 2018), and semantic textual similarity (Hoogeveen et al., 2015; Lei et al., 2016). Litschko et al. (2022) systematic empirical study focused on the suitability of SOTA multilingual encoders for cross-lingual document and sentence retrieval tasks across a number of diverse language pairs. They benchmark the performance in unsupervised ad-hoc (setup with no relevance judgments for IR-specific fine-tuning) and supervised sentence and document-level CLIR. In other words, they profile the suitability of SOTA pre-trained multilingual encoders for different CLIR tasks and diverse language pairs across unsupervised, supervised and transfer setups. They also propose localized relevance matching for document-level CLIR (independently score a query against document). For unsupervised document-level CLIR, they show that pre-trained multilingual encoders on average fail to significantly outperform earlier models based on CLWEs. They also show that the performance of those multilingual encoders crucially depends on how one encodes semantic information with the models (treating them as sentence/document encoders directly versus averaging over constituent words and/or subwords). Multilingual sentence encoders fine-tuned on labeled data from sentence pair tasks like natural language inference or semantic text similarity as well as using parallel sentences on the other hand are shown to substantially outperform general-purpose models in sentence-level CLIR. The second class focuses on training training models with information retrieval objectives but it is not clear how they generalize to new languages. In our work, we investigate ways to further improve the transfer of these off-purpose sentences on top of semantic specialization in a data-efficient manner.

Multilingual Meta-Transfer Learning Meta-learning has gained the attention of the NLP community recently with applications in cross-domain, cross-problem, and cross-lingual transfer learning (Lee et al., 2022). Meta-learning has been leveraged for semantic search-related tasks but only monolingually. Lin and Chen (2020) is the first work of its kind to device a meta-learning algorithm for information retrieval tasks. They lever-

age model-agnostic meta-learner (MAML) to learn an initialization of model parameters for the re-ranker of documents by reformulating the problem as a N-way K-Shot setup where query is a category and the document corresponding to it as a positive example and four documents not related to the query. They show that their approach improves over baselines involving vanilla DSSM and Vector Space Models. They also show that fine-tuning in addition to meta-learning lead to more gains. However, they use meta-learning just at the level of the ranker and not for other components like searcher in which they only use traditional approaches like Match 25 to calculate the relationship between query documents and retrieval documents. It is not clear whether meta-learning can be used more in an end-to-end fashion or to improve other components. Other meta-learning work which focus on the re-ranking component include Laadan et al. (2019); Carvalho et al. (2008) but they all follow a pipelined approach.

Since there is no prior work leveraging meta-learning for cross-lingual or multilingual semantic search, to the best of our knowledge, we describe in this section. The first work of its kind using meta-learning for cross-lingual transfer learning is Gu et al. (2018), which is applied to neural machine translation. They extend MAML(Finn et al., 2017) to transfer from multilingual high-resource language tasks to to low-resource languages. They show the competitive advantages of cross-lingual meta-transfer learning compared to other multilingual baselines. Other applications include speech recognition (Hsu et al., 2020; Winata et al., 2020; Chen et al., 2020; Xiao et al., 2021), Natural Language Inference(XNLI) (Conneau et al., 2018) and Multilingual Question Answering(MLQA) (Lewis et al., 2020) using X-MAML (Nooralahzadeh et al., 2020), task-oriented dialog (Schuster et al., 2019) and TyDiQA (Clark et al., 2020) using X-METRA-ADA (M'hamdi et al., 2021), dependency parsing (Langedijk et al., 2022).

Most recent work adapting meta-learning to applications involving different languages focus on cross-lingual meta-learning. Multilingual meta-learning differs from cross-lingual meta-transfer learning in its support for multiple languages jointly. M'hamdi et al., for example, propose X-METRA-ADA which performs few-shot learning on one single target language at a time and also enable zero-shot learning on target languages not seen during meta-training or meta-adaptation. Their approach shows gains compared to naive fine-tuning in the few-shot more than the zero-shot learning scenario. Tarunesh et al. (2021) propose a meta-learning framework for both multi-task and multilingual transfer leveraging heuristic sampling approaches. They show that a joint approach to multi-task and multilingual learning using meta-learning enables effective sharing of parameters across multiple tasks and multiple languages thus benefits deeper semantic analysis tasks such as QA, PAWS, NLI, etc. van der Heijden et al. (2021) propose a meta-learning framework and show its effectiveness in both the cross-lingual and multilingual training adaptation settings of document classification. However, their multilingual evaluation is focused on the scenario where the same target languages during meta-testing can be also used as auxiliary languages during meta-training. This motivates us to investigate in this paper more in the direction of multilingual meta-transfer learning, where we test the generalizability of our meta-learning model when it is learned by taking into consideration multiple languages jointly for semantic search.

Meta-Distillation Learning Previous works at the intersection of meta-learning and knowledge distillation either use meta-learning as a more effective alternative to the more traditional knowledge distillation methods. Recently, more work has started adopting a meta-learning approach to knowledge distillation by consolidating a feedback loop between the teacher and the student networks where the teacher can learn to better transfer knowledge to the student network (Zhou et al., 2022) or by meta-learning the distillation hyperparameter tuning (Liu et al., 2022). Knowledge distillation has also been leveraged to enhance the portability of MAML networks (Zhang et al., 2020). It has been shown that a portable MAML with a smaller capacity can further boost few-shot learning better than vanilla MAML. To the best of our knowledge, we are the first to explore knowledge distillation to bridge the gap between different cross-lingual meta-transfer learning models and to enhance the alignment between them.

B More Details on Base Models

For asymmetric semantic search, we use a Transformer-based triplet-encoder model. In the original paper on the asymmetric benchmark we evaluate on (Roy et al., 2020), a dual-encoder model is trained using contrastive loss in the form

of an in-batch sampled softmax loss. This format reuses for each question answers from other questions in the same batch (batched randomly) as negative examples. Instead, we use triplet loss (Schroff et al., 2015), which was also shown to outperform contrastive loss in general. Triplet loss is shown to surpass contrastive loss in general.[13] Its strength derives not just from the nature of its function but also from its sampling procedure. This sampling procedure which merely requires positive instances to be closer to negative instances doesn't require gathering as many positive examples as contrastive loss requires. This makes triplet loss more practical in our few-shot learning multilingual/cross-lingual scenario, as it provides more freedom in terms of constructing negative candidates to tweak different sampling techniques from different languages. We thus define a *triplet encoder model* (shown in Figure 7) with three towers encoding the question, its answer combined with its context, and the negative candidates and their contexts. While those towers are encoded separately, they still share the same Transformer encoder model which is initialized with pre-trained Sentence Transformers. On top of that, two dot products $d(q,p)$ and $d(q,n)$ are computed. $d(q,p)$ is the dot product between the question q and its answer p, whereas $d(q,n)$ is between q and its non-answer candidate. Triplet loss is computed as : $\mathcal{L} = \max\left(d(q,p) - d(q,n) + margin, 0\right)$ where $margin$ is a tun-able hyperparameter to eventually make each triplet an easy one by pushing the distance $d(a,p)$ closer to 0 and $d(a,n)$ to $d(a,p) + margin$.

Triplets (q,p,n) can be sampled with different levels of difficulty, as follows:

- **Easy triplets:** $d(q,p) + margin < d(q,n)$.
- **Hard triplets:** $d(q,n) < d(q,p)$.
- **Semi-hard triplets:** $d(q,p) < d(q,n) < d(q,p) + margin$.

For symmetric search, we use a Transformer-based dual-encoder model (shown in Figure 8), which encodes sentence 1 and sentence 2 in each sentence pair separately using the same shared encoder. Then, the cosine similarity score is computed for each sentence pair and the mean squared error (squared L2 norm) is computed between that and the golden score. This is not a retrieval-based task, but a semantic similarity task.

[13] As posited in https://shorturl.at/ktvx9.

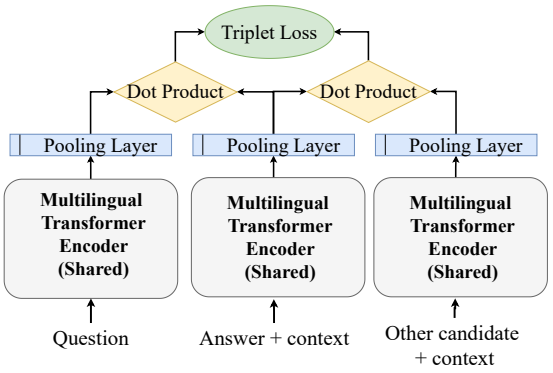


Figure 7: Architecture of Transformer-based triplet encoder for asymmetric semantic search.

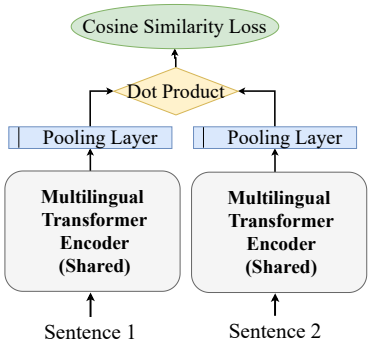


Figure 8: Architecture of Transformer-based dual-encoder for symmetric semantic search.

C More Experimental Setup Details

C.1 Downstream Datasets

Tables 2 and 3 show a summary of the statistics of LAReQA and $\text{STSB}_{\text{Multi}}$ per language and split, respectively. XQuAD-R in LAReQA has been distributed under the CC BY-SA 4.0 license, whereas $\text{STSB}_{\text{Multi}}$ has been released under the Creative Commons Attribution-ShareAlike 4.0 International License. The translated datasets from SQUAD_{EN} and STSB_{EN} are shared under the same license as the original datasets. SQUAD_{EN} is shared under XTREME benchmark Apache License Version 2.0. STSB_{EN} scores are under Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) and sentence pairs are shared under Commons Attribution - Share Alike 4.0 International License).

C.2 Upstream Meta-Tasks

We detail in Table 4 the arrangements of languages for the different meta-tasks used in the meta-training $\mathcal{D}_{\text{meta-train}}$, meta-validation $\mathcal{D}_{\text{meta-valid}}$, and meta-testing $\mathcal{D}_{\text{meta-test}}$ datasets. To make the comparison fair and consistent across different

Language	ISO	Train		Dev		Test	
		#Q	#C	#Q	#C	#Q	#C
Arabic	AR	696	783	220	255	274	184
German	DE	696	812	220	256	274	208
Greek	EL	696	788	220	254	274	192
Hindi	HI	696	808	220	252	274	184
Russian	RU	696	774	220	262	274	183
Thai	TH	696	528	220	178	274	146
Turkish	TR	696	732	220	248	274	187

Table 2: Statistics of LAReQA in each 5-fold cross-validation split. #Q denotes the number of question whereas #C denotes the number of candidates.

Language Pair	ISO	# Sentence Pairs		
		Train	Dev	Test
English-English	EN-EN	150	50	50
Spanish-Spanish	ES-ES	150	50	50
Spanish-English	ES-EN	150	50	50
Arabic-Arabic	AR-AR	150	50	50
Arabic-English	AR-EN	150	50	50
Turkish-English*	TR-EN	150	50	50

Table 3: Statistics of the $\text{STSB}_{\text{Multi}}$ from SEM-Eval2007 in each 5-fold cross-validation split. * means that for Turkish-English, there are only 250 ground truth similarity scores, while there are 500 sentence pairs. We assume that the ground truth scores are only for the first 250 sentence pairs. In addition to that, we use 5749 train, 1500 dev, and 1379 test splits from the STSB original English benchmark.

transfer modes, we use the same combination of languages and tweak them to fit the transfer mode. By picking a high number of meta-tasks during meta-training, meta-validation, and meta-testing, we make sure that all transfer modes are exposed to the same number of questions and candidates. We use Train and Dev splits are used to sample $\mathcal{D}_{\text{meta-train}}$ and $\mathcal{D}_{\text{meta-valid}}$, respectively

C.3 Hyperparameters

Based on our prior investigation of different sentence-transformer models in Table 5, we notice that *paraphrase-multilingual-mpnet-base-v2*[14], which maps sentences and paragraphs to a 768-dimensional dense vector space, performs the best for LAReQA, so we use it in our S-BERT experiments on that dataset. The good initial performance of this pre-trained model is not surprising since it was trained on parallel data and is recommended for use in tasks like clustering or semantic search. For pre-processing LAReQA and SQUAD_{EN}, we truncate/pad all questions to length 96 and all answer or negative candidates concatenated with their contexts to 256. For pre-processing $\text{STSB}_{\text{Multi}}$ and STSB_{EN}, we pad or truncate each sentence to fit the maximum length of 100.

For both benchmarks, for Fine-tune baselines, following XTREME-R, we use AdamW optimizer (Loshchilov and Hutter, 2019). We use a learning rate of $lr = 5e-5$, $\epsilon = 1e-8$ and a weight decay of 0, with no decay on the bias and LayerNorm weights. We use a batch size of 8 triplets or sentence pairs. For LAReQA, we sample 3 negative examples per anchor and then project those to 3 triplets with one negative example and use a margin of 1. In $\text{STSB}_{\text{Multi}}$, we use just sets of sentence pairs composed of one source and one target sentence each, where we don't have negative examples so we don't need to flatten the dimensions of the negative examples. We sample 7,000, 2,000, and 1,000 meta-tasks in the meta-training, meta-validation, and meta-testing phases respectively. We use meta-batches of size 4. In each meta-task, we randomly sample $k = 8$ and $q = 4$ support and query triplets respectively. We use the same meta-tasks and sampling regime in Fine-tune as well.

For MAML and MAML-Align in both benchmarks, we use learn2learn (Arnold et al., 2020) implementation to handle gradient updates, especially in the inner loop. For the inner loop, we use learn2learn pre-built optimizer with a learning rate $\alpha = 1e-3$. The inner loop is repeated $n = 5$ times for meta-training and meta-validation and meta-testing. For the outer loop, we use the same optimizer with the same learning rate $\beta = 1e-5$ that we used in the Fine-tune model. At the end of each epoch, we perform meta-validation similarly to meta-training with the same hyperparameters described before. We use the same hyperparameters for MAML-Align for both T-MAML and S-MAML except that we run the gradient updates in the inner loop in S-MAML just once, whereas for T-MAML we perform $n = 5$ inner loop gradient updates. We jointly optimize the outer loop losses weighting the knowledge distillation by $\lambda = 0.5$. We don't use meta-testing but keep it for evaluation purposes. For a consistent comparison, we don't use meta-testing for our main evaluation as we use standard testing cross-validation splits, but we will include those meta-testing datasets to encourage future work on few-shot learning. All experiments are run for one fixed initialization seed using a 5-fold cross-validation. We observe a variance with respect to different seeds smaller than the variance with respect to 5-fold cross-validation, so we re-

[14] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2.

Transfer Mode	Phase	Support→Query/Support1→Support2→Query	
		LAReQA	$\text{STSB}_{\text{Multi}}$
mono→mono	All	EL_EL→AR_AR HI_HI→DE_DE	(EN_EN,AR_AR,ES_ES)→(EN_EN,AR_AR,ES_ES)
mono→bi	All	EL_EL→EL_AR HI_HI→HI_DE	[EN_EN,AR_AR,ES_ES]→[AR_EN,ES_EN,TR_EN]
mono→multi	All	EL_EL→EL_{AR,EL} HI_HI→HI_{DE,HI}	Not Applicable
bi→multi	All	EL_AR→EL_{AR,EL} HI_DE→HI_{DE,HI}	Not Applicable
mixt	All	mono→mono mono→bi mono→multi bi→multi	Not Applicable
trans	Meta-train Meta-valid	mono→bi bi→multi	Not Applicable
mono→bi→multi	All	EL_EL→EL_AR→EL_{AR,EL,HI} HI_HI→HI_DE→HI_{AR,DE,HI}	EN_EN→AR_EN→EN_{AR,EN,ES} AR_AR→AR_ES→AR_{AR,EN,ES} ES_ES→ES_AR→ES_{AR,EN,ES}

Table 4: Arrangements of languages for the different modes of transfer and meta-learning stages for two standard benchmark datasets LAReQA and $\text{STSB}_{\text{Multi}}$. X→Y denotes transfer from an X model (for example a monolingual model) used to sample the support set to a Y model (for example bilingual model) used to sample the query set. We denote a support or query set in LAReQA by x_y where x and y are the ISO language codes of the question and the candidate answers and x_y in $\text{STSB}_{\text{Multi}}$ where x and y are the ISO language codes of sentence 1 and 2 respectively. We use parenthesis to mean that the same language pairs cannot be used in both support and query sets, brackets to denote non-exclusivity (or in other words the language pairs used as a support can also be used as a query), and curled braces to mean the query set may be sampled from more than one language. We do not experiment with mono→multi, bi→multi, mixt, and trans for $\text{STSB}_{\text{Multi}}$, since it is not a multilingual parallel benchmark, but we still experiment with mono→bi→multi using machine-translated data in that case.

Sentence Transformers Model	mAP@20
LASER	13.5 ± 0.7
LaBSE	48.7 ± 2.6
M-BERT+SQUAD_{EN}	37.9 ± 3.4
distilbert-multilingual-nli-stsb-quora-ranking	44.1 ± 0.9
use-cmlm-multilingual	36.8 ± 2.6
distiluse-base-multilingual-cased-v2	46.9 ± 2.5
paraphrase-multilingual-MiniLM-L12-v2	49.6 ± 2.7
multi-qa-distilbert-dot-v1	6.4 ± 0.3
paraphrase-multilingual-mpnet-base-v2	**57.0** ± 2.9

Table 5: Comparison of mAP@20 multilingual 5-fold cross-validation evaluation of different S-BERT models compared to M-BERT model. Best results are highlighted in **bold**.

port the latter to have a better upper bound of the variance.

All experiments are conducted on the same computing infrastructure using *one* NVIDIA A40 GPU with 46068 MiB memory and *one* TESLA P100-PCIE with 16384 MiB memory of CUDA version 11.6 each. We use Pytorch version 1.11.1, Python version 3.8.13, learn2learn version 0.1.7, Hugging Face transformers version 4.21.3 and Sentence-Transformers 2.2.2. For paraphrase-multilingual-mpnet-base-v2 used in the experiments in the main paper, there are 278,043,648 parameters. For asymmetric and symmetric semantic search benchmarks, there are three and two encoding towers, respectively. Therefore, there are 834,130,944 and 556,087,296 parameters used for asymmetric and symmetric semantic search benchmarks, respectively.

For all experiments and model variants, we train for up to 20 epochs maximum and we implement early stopping, where we run the experiment for as long as there is an improvement on the Dev set performance. After 50 mini meta-task batches of no improvement on the Dev set, the experiment stops running. We use the multilingual performance on the Dev set averaged over all languages of the query set as the early stopping evaluation criteria. Based on this early stopping policy, we report in Table 6 the typical runtime for each upstream model variant and baseline.

Model	Runtime
Fine-tune	2 h 18 min
MAML	3 h 19 min
MAML-Align	19 h 29 min

Table 6: Runtime per model variant excluding evaluation.

D More Results

Tables 7 and 8 show full fine-grained results for all languages and language pairs for both semantic search benchmarks.

Model	Train Language(s) Configuration	Testing Languages							
		Few-Shot Languages				Zero-Shot Languages			
		Arabic AR	German DE	Greek EL	Hindi HI	Russian RU	Thai TH	Turkish TR	Mean
Zero-Shot Baselines									
LASER	-	13.2	15.1	14.6	9.4	14.9	13.0	14.1	13.5
LaBSE	-	44.7	47.9	53.0	53.4	53.1	49.8	48.1	50.0
S-BERT	-	56.3	54.6	58.2	57.2	58.7	60.2	54.1	57.0
+Few-Shot Learning									
S-BERT+Fine-tune	mono→mono	45.9	46.3	47.9	45.4	48.9	49.7	45.1	47.0
	mono→bi	45.8	46.5	48.6	45.0	48.9	49.4	45.0	47.0
	mono→multi	40.4	42.5	43.1	37.8	44.1	44.3	41.1	41.9
	bi→multi	33.8	35.6	35.2	32.4	37.1	37.2	34.4	35.1
	mixt	38.3	39.8	40.7	39.3	41.9	41.7	38.7	40.1
	trans	38.7	39.9	41.8	40.1	42.6	42.6	39.4	40.7
S-BERT+MAML	mono→mono	56.3	54.5	58.5	57.0	59.3	59.6	53.8	57.0
	mono→bi	55.9	55.0	58.4	56.9	58.8	59.9	54.2	57.0
	mono→multi	54.9	53.6	57.0	55.8	57.7	58.7	53.1	55.9
	bi→multi	54.5	53.6	56.6	55.5	57.3	58.5	52.8	55.5
	mixt	55.0	53.9	57.2	55.3	57.6	58.7	52.9	55.8
	trans	56.0	54.8	59.1	57.0	59.1	59.9	54.4	57.2
S-BERT+MAML-Align	mono→bi→multi	*57.0*	*55.1*	*59.2*	*57.7*	*59.5*	*60.2*	*54.6*	*57.6*
+Machine Translation									
S-BERT+T-Train+Fine-tune	AR_AR→AR_AR	46.6	45.8	48.8	46.8	49.3	48.6	44.9	47.3
	DE_DE→DE_DE	45.9	45.1	48.2	45.8	49.0	48.8	44.5	46.8
	EL_EL→EL_EL	43.5	43.1	43.8	43.4	46.5	45.0	41.7	43.8
	HI_HI→HI_HI	46.5	44.8	47.1	45.9	48.4	49.6	43.7	46.6
	All test languages	44.8	43.5	46.9	44.0	47.0	46.4	42.1	45.0
S-BERT+T-Train+MAML	AR_AR→AR_AR	**57.3**	**55.3**	**59.3**	**58.3**	**60.2**	**60.7**	**54.8**	**58.0**
	DE_DE→DE_DE	56.1	54.4	58.3	57.1	58.8	59.8	54.1	56.9
	EL_EL→EL_EL	55.9	53.1	57.4	56.3	58.5	59.2	52.8	56.2
	HI_HI→HI_HI	56.7	54.0	58.5	57.1	58.9	60.3	53.7	57.0
	All test languages	55.9	53.8	58.0	56.6	58.1	59.2	53.4	56.4

Table 7: mAP@20 multilingual 5-fold cross-validated performance tested for different languages. Best and second-best results for each language are highlighted in **bold** and *italicized* respectively, whereas best results across categories of models are underlined. Gains from meta-learning approaches are consistent across few-shot and zero-shot languages.

Model	Train Language(s) Configuration	Testing Languages						
		Arabic-Arabic AR-AR	Arabic-English AR-EN	Spanish-Spanish ES-ES	Spanish-English ES-EN	English-English EN-EN	Turkish-English TR-EN	Mean
Zero-Shot Learning								
LASER	-	22.5 ± 8.5	21.6	33.1	15.3	31.1	21.2	24.1
LaBSE	-	71.6	73.2	83.2	68.7	76.3	74.9	74.6
S-BERT	-	77.6	81.3	84.6	83.7	85.5	75.7	81.4
+Few-Shot learning								
S-BERT+Fine-tune	mono→bi	77.2	77.8	86.2	79.6	85.0	73.7	79.9
S-BERT+MAML	mono→bi	77.6	80.9	85.1	83.5	85.6	75.5	81.3
S-BERT+MAML-Align	mono→bi→multi	**79.0**	80.6	86.6	81.5	**90.6**	**76.3**	**82.4**
+Machine Translation								
S-BERT+T-Train+Fine-tune	AR_AR→AR_AR	59.5	50.6	82.7	70.1	82.4	62.5	68.0
	EN_EN→EN_EN	72.6	73.1	82.4	72.2	80.3	68.8	74.9
	ES_ES→ES_ES	74.2	72.3	82.3	66.8	79.7	68.5	73.9
	TR_TR→TR_TR	73.9	74.6	85.9	79.6	84.3	68.5	77.8
	All test languages	65.8	63.0	82.5	75.8	83.0	67.8	73.0
S-BERT+T-Train+MAML	AR_AR→AR_AR	75.5	80.5	85.8	83.1	85.6	75.0	80.9
	EN_EN→EN_EN	*77.8*	*81.7*	85.1	**83.8**	*85.7*	75.8	*81.6*
	ES_ES→ES_ES	76.4	79.4	*86.9*	80.4	84.7	74.1	80.3
	TR_TR→TR_TR	77.2	79.8	**87.3**	81.6	84.5	74.2	80.8
	All test languages	77.6	**81.8**	84.7	83.6	85.6	*75.9*	81.5

Table 8: Pearson correlation Pearson's r × 100 5-fold cross-validated performance on $\text{STSB}_{\text{Multi}}$ benchmark using different models few-shot learned on $\text{STSB}_{\text{Multi}}$ or its translation. Best and second-best results for each language are highlighted in **bold** and *italicized* respectively, whereas best results across categories of models are underlined.