Cross-Lingual Knowledge Distillation for Answer Sentence Selection in Low-Resource Languages

Shivanshu Gupta *
University of California, Irvine
shivag5@uci.edu

Yoshitomo Matsubara
Amazon Alexa AI
yomtsub@amazon.com

Ankit Chadha
Amazon Alexa AI
ankitrc@amazon.com

Alessandro Moschitti
Amazon Alexa AI
amosch@amazon.com

Abstract

While impressive performance has been achieved on the task of Answer Sentence Selection (AS2) for English, the same does not hold for languages that lack large labeled datasets. In this work, we propose Cross-Lingual Knowledge Distillation (CLKD) from a strong English AS2 teacher as a method to train AS2 models for low-resource languages in the tasks without the need of labeled data for the target language. To evaluate our method, we introduce 1) Xtr-WikiQA,[1] a translation-based WikiQA dataset for 9 additional languages, and 2) TyDi-AS2,[2] a multilingual AS2 dataset with over 70K questions spanning 8 typologically diverse languages. We conduct extensive experiments on Xtr-WikiQA and TyDi-AS2 with multiple teachers, diverse monolingual and multilingual pretrained language models (PLMs) as students, and both monolingual and multilingual training. The results demonstrate that CLKD either outperforms or rivals even supervised fine-tuning with the same amount of labeled data and a combination of machine translation and the teacher model. Our method can potentially enable stronger AS2 models for low-resource languages, while TyDi-AS2 can serve as the largest multilingual AS2 dataset for further studies in the research community.

1 Introduction

Answer Sentence Selection (AS2) is the task of ranking a given set of answer candidates according to their probability of correctly answering a given question. This is a core task for retrieval-based web Question Answering (QA) systems. Indeed, AS2 models applied to the sentences of documents relevant to a question, *e.g.*, retrieved by a search engine, provide accurate answers.

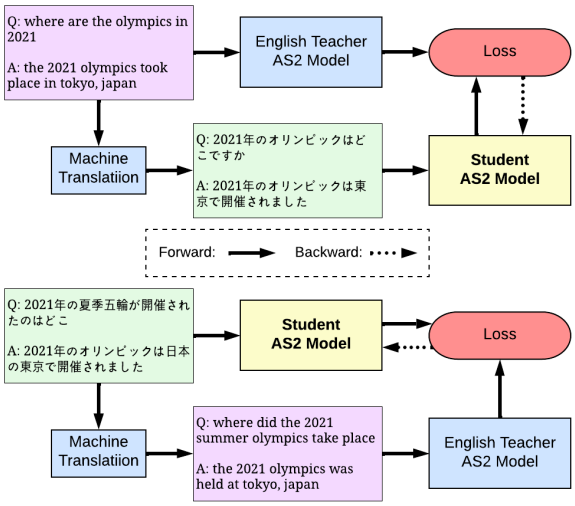

Figure 1: Cross-Lingual Knowledge Distillation (CLKD) in two different scenarios: (**Top**) using unlabeled English AS2 dataset for target low-resource language lacking any data and (**Bottom**) using unlabeled original low-resource language AS2 dataset. CLKD enables student AS2 models to learn from English teacher AS2 models without human-annotated datasets.

While AS2 has been extensively studied for English (Wang and Nyberg, 2015; Chen et al., 2017; Tan et al., 2017; Tymoshenko and Moschitti, 2018; Nicosia and Moschitti, 2018; Garg et al., 2020; Tian et al., 2020; Matsubara et al., 2020; Laskar et al., 2020; Bonadiman and Moschitti, 2020; Soldaini and Moschitti, 2020; Lauriola and Moschitti, 2021; Krishnamurthy et al., 2021; Han et al., 2021; Zhang et al., 2021; Mrini et al., 2021; Di Liello et al., 2022; Matsubara et al., 2022), much less research has been devoted to other languages. This is despite the rapidly increasing importance of multilingual QA with the proliferation of conversational agents and voice assistants using multilingual content from the Web to target locales across the world (Li et al., 2022). A major barrier to achieving similar performance obtained with English in other languages is the lack of large labeled datasets. However, labeling AS2 datasets for every language will be

*This work was done as an intern at Amazon Alexa AI.

[1] https://huggingface.co/datasets/AmazonScience/xtr-wiki_qa

[2] https://huggingface.co/datasets/AmazonScience/tydi-as2

Findings of the Association for Computational Linguistics: ACL 2023, pages 14078–14092

prohibitively expensive as even a single AS2 instance can contain hundreds of candidate answers per question (see Table 2). This necessitates methods that do not require labeled target language data. A simple approach is to just translate questions to English and then use an English AS2 model (Vu and Moschitti, 2021; Asai et al., 2021). While this pipeline can be quite accurate (Li et al., 2022), the need for machine translation makes inference slow and inefficient. An alternative approach would be to train AS2 models on target language translations of English datasets. However, training using *translationese* seems sub-optimal due to errors and artifacts introduced by machine translation. Moreover, models trained on English questions could be ill-suited for answering target language questions due to information asymmetry (Asai et al., 2021) *i.e.*, questions asked in English are likely to differ from those in the target language due to cultural bias, *e.g.*, they can refer to different entities.

In this work, we propose Cross-Lingual Knowledge Distillation (CLKD) as a method to use readily available and highly-accurate English AS2 models to train AS2 models for low-resource languages lacking labeled data. CLKD can use English datasets to train AS2 models for languages lacking any data and can further leverage unlabeled original target language data without the need for costly manual annotation.

Figure 1 illustrates a high-level description of our approach. CLKD works similarly to classic Knowledge Distillation (KD) (Hinton et al., 2015) in that a student model is trained to mimic a teacher. The main novelty of our approach is the fact that the teacher and student models operate in different languages, namely the source and target languages. Thus, the input question-answer pairs are translated into both the source and the target languages. Additionally, to allow use of original target language data, which is typically unlabeled, we use only the soft labels obtained from the teacher even when gold labels are available, *i.e.*, given an unlabeled input question and candidate answer pair (q, a), the student is trained using the Kullback-Leibler divergence loss between the probability scores of the teacher and the student when applied to (q, a).

To evaluate our approach for diverse languages, we construct two new multilingual AS2 datasets: Xtr-WikiQA[3] (10 languages) and TyDi-AS2 (8 languages). Xtr-WikiQA consists of 10-language parallel corpora, including the original English corpus from WikiQA (Yang et al., 2015).[4] To enable evaluation with original target language data, we further create TyDi-AS2 by converting TyDi-QA (Clark et al., 2020a),[5] a multilingual QA dataset to an AS2 dataset. We also translate the English TyDi-AS2 dataset to all the other TyDi-AS2 languages to build an additional translationese corpus.

Using the above datasets, we perform extensive experiments with multiple teacher models, about 20 different student models of varying sizes, including both monolingual and multilingual PLMs, and both monolingual and multilingual training. Additionally, to evaluate the utility of our method for both languages lacking any data and for those with some unlabeled data, we experiment with both using only English data (Xtr-WikiQA and English TyDi-AS2) and only unlabeled target language data (TyDi-AS2). We show that CLKD consistently either rivals or outperforms even supervised fine-tuning with the same amount of gold-labeled data demonstrating the benefit of CLKD using soft labels obtained from a strong English AS2 teacher model. In particular, we show that CLKD using original language unlabeled data outperforms 1) fine-tuning with gold-labeled translationese and 2) for larger students, even the MT+English AS2 model pipeline, demonstrating the importance of original target language data.

We expect that the ability of CLKD to train AS2 models without the need for costly annotation process will enable stronger AS2 performance for the world's many low-resource languages. To support further studies on AS2 tasks for such languages, we will make the datasets introduced in this work and our trained models publicly available.

2 Related Work

We briefly summarize the related studies.

2.1 KD for Model Compression

KD was originally proposed as a method for model compression that improves the performance of a weaker model to be trained (student) by learning from a strong but cumbersome model (teacher) (Hinton et al., 2015). With large pretrained language models based on Transformer (Vaswani et al., 2017) becoming the new

[3]Xtr-WikiQA: **X-tr**anslated-**WikiQA**

[4]https://www.microsoft.com/en-us/download/details.aspx?id=52419

[5]https://ai.google.com/research/tydiqa/

paradigm for natural language processing (NLP) tasks, KD has gained greater attention from the NLP community, with many studies on KD for Transformer-based models (Sanh et al., 2019; Jiao et al., 2020; Lu et al., 2020; Park et al., 2021a). For AS2 tasks, Matsubara et al. (2022) propose a multi-head student model (CERBERUS) to distill knowledge in an ensemble of multiple diverse teacher models to improve model accuracy without significantly increasing model complexity.

2.2 Learning from Teacher Models in Different Domains/Tasks

Garg and Moschitti (2021) propose a technique to filter out non-answerable questions in question-answering systems, which trains a binary classifier for an input question text by mimicking the confidence score from a pretrained AS2 model (input: pair of question and candidate answer). Gabburo et al. (2022) leverage an AS2 model (a discriminative ranking model) as a teacher model to train an answer generation model (Hsu et al., 2021).

There are also a few related studies regarding KD in cross-lingual problem settings. To address the lack of Chinese sentiment corpora, Wan (2009) leverages machine translations (English-to-Chinese and Chinese-to-English) and studies a cross-lingual sentiment classification problem. Xu and Yang (2017) also work on sentiment analysis tasks and propose a cross-language distillation with feature adaptation. Reimers and Gurevych (2020) propose a method to extend existing (English) sentence embedding models to new languages for multilingual student models. Karamanolakis et al. (2020) present a text classification model training method with a small budget of word-level translations for words that are most indicative of the target task and unlabeled documents in the target language.

Li et al. (2022) propose a multi-stage KD to learn a cross-lingual document retriever from an English retriever, which is the most relevant work to ours. While similar in regard to learning from an English model, our approach significantly differs from theirs. First, Li et al. (2022) train cross-lingual document retrievers (*i.e.*, query and document differ in language), whereas we focus on AS2 models taking as input question and answer in the same language, while the teacher is in English and the student is in another language. Second, while their multi-stage KD method requires the student and teacher to share the embedding size, our single-stage KD method does not have this restriction. Finally, they evaluate their method on XLM-RoBERTa (Conneau et al., 2020) only (as student and teacher models), whereas we perform a much more comprehensive study spanning multiple teachers and approximately 20 different pretrained language models.

3 Knowledge Distillation for AS2

3.1 AS2 Task

We consider the task of Answer Sentence Selection where given a question q and a set of answer sentence candidates, $S = \{s_1, \ldots, s_n\}$, the goal is to select the sentence s^* that best answers the question. Following prior work (Garg et al., 2020), we frame this as a ranking task where we assign a score to each sentence s_i for the question q and then select the sentence with the highest score. Formally, given a question-sentence pair (q, s), the AS2 model $\mathcal{M}$ produces a score $\mathcal{M}(q, s)$ measuring the likelihood of s being the correct answer to q. We then select the sentence with the highest score as the answer, *i.e.*, $s^* = \text{argmax}_{s \in S}\, \mathcal{M}(q, s)$.

3.2 Knowledge Distillation

Knowledge Distillation (Hinton et al., 2015) is an effective method to transfer knowledge from a strong teacher model T to a student model $\mathcal{S}$, by training the student to mimic the teacher. Formally, given inputs $\{x_i\}_{i=1}^{N}$, the distillation loss is a weighted sum of cross-entropy ($\mathcal{L}_{\text{CE}}$) of the student w.r.t. gold labels and KL-divergence ($\mathcal{L}_{\text{KL}}$) of the teacher and student's class probabilities,

$$\begin{aligned}\mathcal{L}_{\text{KD}}(x, y) =& \alpha \mathcal{L}_{\text{CE}} \left(\text{softmax}\left(\mathbf{z}^{(\mathcal{S})}\right), y\right) + \\ & (1-\alpha)\tau^2 \mathcal{L}_{\text{KL}}(\mathbf{p}_i^{(T)}, \mathbf{p}_i^{(\mathcal{S})}), \qquad (1) \\ \mathbf{p}_i^{(T)} =& \text{softmax}\left(\mathbf{z}^{(T)}/\tau\right) \\ \mathbf{p}_i^{(\mathcal{S})} =& \text{softmax}\left(\mathbf{z}^{(\mathcal{S})}/\tau\right),\end{aligned}$$

where $\mathbf{z}_i^{(T)} = T(x_i)$ and $\mathbf{z}_i^{(\mathcal{S})} = \mathcal{S}(x_i)$ are logits from teacher and student models, respectively. y indicates a gold label (human annotation), and α and τ ("temperature") are hyperparameters.

3.3 Cross-Lingual Knowledge Distillation

In order to train AS2 models for low-resource languages lacking labeled data, we propose Cross-Lingual Knowledge Distillation (CLKD) from accurate and readily available English AS2 models.

CLKD assumes the absence of gold labels for target languages and, in general, teaches the student model for a "target" language to mimic the teacher from a different "source" language (English in this study) as illustrated in Fig. 1. In other words, the CLKD loss is the second term of Eq. 1 ($\alpha = 0$, no gold labels are used) with student and teacher logits obtained by feeding them the same input in target and source language, respectively.

Formally, given a teacher T^l for source language l, and two parallel unlabeled datasets, $D^{(l)} = \left\{x_i^{(l)}\right\}_{i=1}^{N}$ and $D^{(l')} = \left\{x_i^{(l')}\right\}_{i=1}^{N}$, for source and target languages, l and l', respectively, CLKD trains a student model $\mathcal{S}^{l'}$ using the same loss as the monolingual distillation case (Eq. 1) but with the teacher and student logits obtained as $\mathbf{z}_i^{(T)} = T(x_i^{(l)})$ and $\mathbf{z}_i^{(\mathcal{S})} = \mathcal{S}(x_i^{(l')})$, respectively.

For the AS2 task, the input will be question-sentence pairs *i.e.*, $x_i = (q_i, s_i)$. Additionally, as we distill knowledge in English AS2 models, the source and target languages will be English and a low-resource language, respectively. Also, since parallel datasets are likely not available, they will be obtained using automatic machine translation.

Depending on the low-resource language data available, CLKD can be applied in two different ways: (1) In absence of any target-language data, CLKD can be applied using an English AS2 dataset (see Fig. 1 top). In this scenario, the teacher and student will be fed the original English and translationese instances, respectively. While this can be applied to any language, errors and artifacts inevitably introduced by machine translation and information asymmetry due to cultural bias with respect to the target language (Asai et al., 2021) will limit the student's performance. (2) CLKD allows overcoming this limitation by utilising original target language unlabeled data when available. As shown in the Fig. 1 bottom, this would involve feeding the original language and English-translated input to the student and teacher models, respectively.

Note that the success of CLKD, particularly with original data, relies on two practical assumptions: (i) two AS2 models for two different languages should produce similar probability scores when applied to inputs that are translations of each other, and (ii) the teacher working on automatically translated data is still accurate enough to transfer *useful* knowledge to the student.

	train	dev	test
#Queries	873	126	243
#QA pairs	8,671	1,130	2,351
#Correct answers	1,040	140	293

Table 1: Statistics of Xtr-WikiQA for each language.

4 New Datasets

In this section, we introduce two new AS2 datasets, Xtr-WikiQA and TyDi-AS2. The datasets are constructed from the WikiQA (Yang et al., 2015) and TyDi-QA (Clark et al., 2020a) datasets, respectively, and the intended use of our new datasets follows Community Data License Agreement (CDLA) - Permissive (Version 2.0).[6]

4.1 Xtr-WikiQA

WikiQA (Yang et al., 2015) has been used as an English AS2 dataset in various studies on AS2 tasks (Garg et al., 2020; Matsubara et al., 2020; Lauriola and Moschitti, 2021). Following (Garg et al., 2020; Matsubara et al., 2022), we remove queries which have no correct answers from the training split, but leave such queries in the development and test splits.

We translate WikiQA using Amazon Translate[7] to construct a new multilingual AS2 dataset, named Xtr-WikiQA,[3] comprising 9 additional languages (*i.e.*, 10 languages in total): Arabic (ar), Dutch (nl), French (fr), German (de), Hindi (hi), Italian (it), Japanese (ja), Portuguese (pt), and Spanish (es). Table 1 shows the statistics of Xtr-WikiQA dataset. Each of the 10 language corpora in Xtr-WikiQA has the same statistics as those are parallel corpora.

4.2 TyDi-AS2

In addition to our *translationese dataset* above, we need a large and accurate multilingual AS2 dataset to evaluate our method and compare against supervised baselines on original target language data. Due to the lack of such datasets, we introduce TyDi-AS2, a large multilingual AS2 benchmark derived from the TyDi-QA dataset (Clark et al., 2020a), a multilingual Machine Reading dataset. TyDi-AS2 is a collection of AS2 datasets for eight typologically diverse languages, including Bengali (bn), English (en), Finnish (fi), Indonesian (id), Japanese (ja), Korean (ko), Russian (ru), and Swahili (sw). The dataset was constructed from the data for the

[6] https://cdla.dev/permissive-2-0/
[7] https://aws.amazon.com/translate/

Language	#Queries			#Sentences			Avg. Sentence Length			#Positive QA Pairs		
	train	dev	test	train	dev	test	train	dev	test	train	dev	test
Bengali (bn)	7,978	2,056	316	1,376,432	351,186	37,465	106.3	106.3	106.7	1,914	472	148
English (en)	6,730	1,686	918	1,643,702	420,899	249,513	107.8	106.6	107.4	2,953	699	810
Finnish (fi)	10,859	2,731	1,870	1,567,695	408,205	298,093	123.3	122.4	123.5	5,317	1,316	1,211
Indonesian (id)	9,310	2,339	1,355	960,270	236,076	97,057	154.6	155.3	153.9	2,237	608	408
Japanese (ja)	11,848	2,981	1,504	3,183,037	822,654	444,106	45.2	45.2	46.1	3,513	846	858
Korean (ko)	7,354	1,943	1,389	1,558,191	392,361	199,043	84.2	84.2	88.4	586	141	216
Russian (ru)	9,187	2,294	1,395	3,190,650	820,668	367,595	109.3	110.0	101.6	5,101	1,277	1,039
Swahili (sw)	8,350	2,850	1,896	1,048,303	269,894	74,775	145.3	144.0	141.0	976	244	356

Table 2: Statistics of TyDi-AS2.

primary task in TyDi-QA, where each instance is accompanied by a Wikipedia article.

Conversion TyDi-QA is a QA dataset spanning questions from 11 typologically diverse languages. Each instance comprises a human-generated question, a single Wikipedia document as context, and one or more spans from the document containing the answer. To convert each instance into AS2 instances, we split the context document into sentences and use the answer spans to identify the correct answer sentences.

To split documents, we use multiple different sentence tokenizers for the diverse languages and omit languages for which we could not find a suitable sentence tokenizer: 1) `bltk`[8] for Bengali, 2) `blingfire`[9] for Swahili, Indonesian, and Korean, 3) `pysdb`[10] (Sadvilkar and Neumann, 2020) for English and Russian, 4) `nltk`[11] (Bird et al., 2009) for Finnish, and 5) `Konoha`[12] for Japanese.

Translation For CLKD experiments with original target language data, we use Amazon Translate[7] to translate the non-English corpora of TyDi-AS2 datasets into English. Furthermore, to conduct another translationese experiments, we also translate the English TyDi-AS2 dataset to all the other TyDi-AS2 languages, similar to (Vu and Moschitti, 2021). We refer to this dataset as Xtr-TyDi-AS2.

Dataset Statistics As the original TyDi-QA test set is not publicly available, we repurposed the dev set for test set and used an 80-20 split of the original training set to create TyDi-AS2's training and dev sets. Table 2 shows statistics of TyDi-AS2.

[8] https://github.com/saimoncse19/bltk

[9] https://github.com/microsoft/BlingFire

[10] https://github.com/nipunsadvilkar/pySBD

[11] https://www.nltk.org/

[12] https://github.com/himkt/konoha

5 Experimental Setup

For a rigorous assessment of the efficacy of CLKD, we design various experiments with different teachers, students, and training data.

5.1 AS2 Models

This section describes Transformer (Vaswani et al., 2017) models we use as AS2 models. Table 3 shows the full list of Hugging Face Transformers (Wolf et al., 2020) pretrained language models used in this study.

Table 3 summarizes the pretrained language models used in this study. We note that the teacher models in Table 3 are fine-tuned on the original English corpus in the target datasets, thus there are two ELECTRA-Large models separately fine-tuned to be teachers for Xtr-WikiQA and TyDi-AS2.

5.1.1 English Teacher Models

To ensure non-specificity to a particular teacher, we experiment with two English AS2 models as teachers in CLKD for Xtr-WikiQA: RoBERTa-Large (Liu et al., 2019) and ELECTRA-Large (Clark et al., 2020b) are the teacher models trained by (Matsubara et al., 2022) for WikiQA using TANDA, a state-of-the-art AS2 model training method (Garg et al., 2020). For TyDi-AS2, we use the ELECTRA-Large model fine-tuned by TANDA on the TyDi-AS2 English dataset instead of WikiQA as the teacher.

5.1.2 Student Models

We experiment with both monolingual and multilingual pretrained language models (PLMs) as students. Additionally, while we experiment with monolingual training for both the two types of student PLMs, for multilingual students, we also experiment with multilingual training using data for all the languages in the corresponding dataset.

Language	Dataset	Hugging Face Pretrained Model	Size	Note
en	Xtr-WikiQA	roberta-large	355M	(Liu et al., 2019) Fine-tuned by Garg et al. (2020)
en	Xtr-WikiQA TyDi-AS2	google/electra-large-discriminator	335M	(Clark et al., 2020b) Fine-tuned by Matsubara et al. (2022)
ar	Xtr-WikiQA	asafaya/bert-base-arabic	111M	(Safaya et al., 2020)
de	Xtr-WikiQA	bert-base-german-cased	109M	
hi	Xtr-WikiQA	monsoon-nlp/hindi-bert	14.7M	
it	Xtr-WikiQA	dbmdz/bert-base-italian-xxl-cased	111M	
ja	Xtr-WikiQA TyDi-AS2	nlp-waseda/roberta-base-japanese	111M	
nl	Xtr-WikiQA	GroNLP/bert-base-dutch-cased	109M	(de Vries et al., 2019)
pt	Xtr-WikiQA	neuralmind/bert-base-portuguese-cased	109M	(Souza et al., 2020)
bn	TyDi-AS2	csebuetnlp/banglabert	111M	(Bhattacharjee et al., 2022)
fi	TyDi-AS2	TurkuNLP/bert-base-finnish-cased-v1	125M	(Virtanen et al., 2019)
id	TyDi-AS2	indobenchmark/indobert-base-p1	124M	(Wilie et al., 2020)
ko	TyDi-AS2	klue/bert-base	111M	(Park et al., 2021b)
ru	TyDi-AS2	DeepPavlov/rubert-base-cased	178M	(Kuratov and Arkhipov, 2019)
sw	TyDi-AS2	Davlan/bert-base-multilingual-cased-finetuned-swahili	178M	Fine-tuned on Swahili corpus
multi	Xtr-WikiQA TyDi-AS2	bert-base-multilingual-cased	178M	(Devlin et al., 2019)
multi	Xtr-WikiQA TyDi-AS2	xlm-roberta-base	278M	(Conneau et al., 2020)
multi	Xtr-WikiQA TyDi-AS2	xlm-roberta-large	560M	(Conneau et al., 2020)

Table 3: List of pretrained language models used in this study.

Monolingual Student Models For experiments with Xtr-WikiQA, we use ELECTRA-Base (Clark et al., 2020b) pretrained on Hindi corpus and BERT-Base (Devlin et al., 2019) pretrained on Arabic, German, Italian, Japanese, Dutch, and Portuguese, respectively. We did not find working monolingual PLMs for Spanish and French. For TyDi-AS2, we use ELECTRA-Base (Clark et al., 2020b) pretrained on Bengali corpus, mBERT (Devlin et al., 2019) finetuned on Swahili corpus, and BERT-Base pretrained on Finnish, Indonesian, Japanese, Korean, and Russian respectively. Table 3 includes the pretrained monolingual models used in this study.

Multilingual Student Models As pretrained multilingual student models, we use mBERT (Devlin et al., 2019), XLM-RoBERTa-Base (XLM-R-Base), and XLM-RoBERTa-Large (XLM-R-Large) (Conneau et al., 2020).

5.2 Training Languages

In addition to pretrained monolingual and multilingual student models, we also experiment with mono- and multilingual training. For monolingual training, we train the model using a single language's training data. We refer to this setting as SINGLE. For the multilingual setting, which is only possible for multilingual models, we use data for all the languages in a particular dataset, which we refer to as ALL.

5.3 Methods

For each dataset, student model, and training languages (SINGLE or ALL), we compare two approaches: direct finetuning using gold labels and CLKD using a teacher's soft labels. We refer to these as FINETUNE and CLKD, respectively. In particular, we use CLKD[E] and CLKD[R] to denote CLKD with ELECTRA-Large and RoBERTa-Large as English teachers, respectively. Finally, for experiments using original language data (*i.e.*, with TyDi-AS2), we additionally compare with the MT-English AS2 pipeline, which involves directly feeding the English translations of the test instances to the English Teacher. This is considered as a strong baseline in (Asai et al., 2021) and (Li et al., 2022).

Note that a potential baseline could be to translate all the English data that the teacher was trained on to the target language. However, this is not a feasible approach as (i) the data may not be available, and (ii) even if it were, it would be prohibitively expensive to translate and retrain for every language. Moreover, it will still suffer from the shortcomings of training on translationese such as artifacts from MT and cultural bias as described in § 3.3.

Language	Student LM	Method	ar	de	es	fr	hi	it	ja	nl	pt
SINGLE	Monolingual	FINETUNE	56.2	63.4	N/A	N/A	27.2	63.8	57.8	59.8	63.7
		CLKD[E]	63.9	70.0	N/A	N/A	27.3	66.0	58.0	65.4	68.6
		CLKD[R]	65.7	72.0	N/A	N/A	28.4	70.1	57.6	66.7	68.6
	mBERT	FINETUNE	61.9	66.0	68.0	63.8	61.3	60.8	66.1	63.5	61.7
		CLKD[E]	69.1	72.6	75.0	71.2	65.6	73.4	71.6	71.5	74.9
		CLKD[R]	69.3	73.1	75.2	71.5	68.3	75.7	71.7	74.9	74.2
	XLM-R-Base	FINETUNE	56.5	56.5	59.9	57.9	56.2	58.2	59.0	31.3	56.8
		CLKD[E]	66.3	64.6	71.5	65.2	64.1	68.2	66.0	65.8	70.0
		CLKD[R]	67.5	64.3	69.1	66.8	65.8	68.5	68.9	58.7	70.0
	XLM-R-Large	FINETUNE	64.2	74.1	71.1	69.1	71.6	66.9	71.3	71.1	75.3
		CLKD[E]	76.3	81.5	81.3	80.0	79.6	80.9	79.6	81.5	82.7
		CLKD[R]	76.3	81.9	81.6	81.9	80.3	80.7	80.4	80.7	81.9
ALL	mBERT	FINETUNE	70.8	70.2	74.1	69.3	67.8	71.6	69.3	70.4	74.2
		CLKD[E]	74.6	77.1	79.3	76.3	71.1	78.7	74.9	75.5	80.0
		CLKD[R]	76.3	77.9	80.8	75.5	71.3	80.0	76.7	77.6	80.4
	XLM-R-Base	FINETUNE	59.1	63.0	66.5	63.9	61.9	64.6	61.6	63.1	64.3
		CLKD[E]	71.6	72.6	73.4	70.9	70.7	74.5	68.7	69.8	75.0
		CLKD[R]	73.8	73.9	74.1	70.1	70.2	73.5	69.3	71.3	75.0
	XLM-R-Large	FINETUNE	75.6	82.9	78.2	80.0	78.9	79.4	78.5	80.4	81.8
		CLKD[E]	80.3	82.7	81.9	82.5	81.3	**83.8**	**82.5**	**83.5**	85.0
		CLKD[R]	**81.6**	**82.6**	**84.2**	**84.5**	**80.4**	82.9	81.8	82.0	**85.1**

Table 4: **Xtr-WikiQA**: Averaged test results (P@1) of models trained on dataset of 1) single target language and 2) all the target languages. We highlight better results(FINETUNE vs. CLKD) and additionally use a bold font for **the best student model** for each of the nine target languages. Our English teacher models, ELECTRA-Large (E) and RoBERTa-Large (R), achieved 87.7% and 91.8% P@1 respectively.

5.4 Training Details

For every model and training configuration, we run three training sessions with different random seeds and present average results. Our implementation is based on PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020) with Python 3.7.

Unless specified otherwise, we use the same hyperparameters for both the supervised baselines and CLKD. To train AS2 models, we use AdamW (Loshchilov and Hutter, 2019) with an initial learning rate of 10^{-6} and the linear scheduler with a warm-up for the first 2.5% of the training iterations. The number of training iterations (model updates) is set to 20,000 and 40,000 for Xtr-WikiQA and TyDi-AS2, respectively. For better training convergence with multilingual training, we increase the number of training iterations to 150,000. We use only 1 GPU to train each of the AS2 models.

In this study, we select $\tau \in \{1, 3, 5, 7\}$ based on the results for the development split[13] and report the averaged test results of the selected AS2 model individually trained with three different random seeds. To run the extensive amount of experiments, we use Amazon EC2[14] instances of `p2.8xlarge`, `p3.8xlarge`, and `p3dn.24xlarge`.

6 Evaluations

We now describe the results of our experiments. In § 6.1, we have a problem setting where we assume that no target language data is available for training and we use translations of the English data instead. In § 6.2, we use the TyDi-AS2 dataset to experiment with the setting where some original target language unlabeled data is available.

6.1 Translationese

Tables 4 and 5 show the results for all the experiments with the Xtr-WikiQA and Xtr-TyDi-AS2 translationese datasets, respectively. Note that the experiments with the Xtr-TyDi-AS2 dataset use the test split of TyDi-AS2 for the evaluation. It is

[13] Tables 7- 9 in Appendix B show the selected temperatures.

[14] https://aws.amazon.com/ec2/

Language	Student LM	Method	bn	fi	id	ja	ko	ru	sw
	MT + TEACHER		63.9	69.2	81.0	55.4	77.8	66.8	86.4
SINGLE	Monolingual	FINETUNE	34.6	54.6	66.1	24.6	71.6	53.7	66.9
		CLKD[E]	35.6	59.7	70.6	26.8	74.7	59.0	69.7
	mBERT	FINETUNE	37.4	52.4	69.7	31.6	70.9	51.7	72.7
		CLKD[E]	43.3	58.6	74.1	35.4	79.5	55.6	74.8
	XLM-R-Base	FINETUNE	29.0	48.8	69.7	33.0	64.9	47.3	67.0
		CLKD[E]	34.4	53.9	72.1	37.8	69.3	51.3	70.1
	XLM-R-Large	FINETUNE	54.9	61.3	77.3	52.0	75.7	59.2	85.3
		CLKD[E]	61.3	65.0	80.9	55.8	78.2	62.2	85.0
ALL	mBERT	FINETUNE	51.0	57.0	74.8	46.3	73.9	56.3	76.9
		CLKD[E]	54.9	61.8	77.4	49.6	80.3	60.3	78.8
	XLM-R-Base	FINETUNE	43.1	53.9	72.6	40.4	70.4	52.5	73.0
		CLKD[E]	50.5	57.9	75.6	44.8	77.2	56.4	77.1
	XLM-R-Large	FINETUNE	62.8	63.9	80.7	57.5	77.3	62.3	83.9
		CLKD[E]	**67.2**	**67.5**	**81.8**	**58.4**	**81.1**	**66.7**	**85.5**

Table 5: **Xtr-TyDi-AS2**: Averaged test results (P@1) of models trained on translationese data for 1) single target language and 2) all the target languages. We highlight better results (FINETUNE vs. CLKD) and additionally use a bold font for **the best student model** for each of the seven target languages.

clear that the performance improves with increasing student model size and going from monolingual training to multilingual training in all languages, even though the training datasets for the diverse languages are translationese from the English corpus. Nevertheless, CLKD consistently outperforms supervised finetuning with gold labels for all the target languages in both the datasets, and we confirm that CLKD significantly improves FINETUNE on nearly all the considered configurations of teacher, student, and both monolingual and multilingual training. This is true even for the Xtr-TyDi-AS2 dataset, which is nearly eight times larger than Xtr-WikiQA, making it even more challenging to reach the performance of the supervised baseline (FINETUNE) with an unsupervised method.

Finally, the performance improvement is greater for smaller models and for monolingual training. This is expected as the teachers have also been trained on the source English dataset, and their performances can be seen as upper bounds for the student performances. The results demonstrate the benefits of soft labels from a strong English AS2 teacher when training AS2 models with no original language data.

6.2 Original Language Data

Table 6 shows results of experiments with the original language datasets in TyDi-AS2. We observe similar trends as with translationese; the performance improves with bigger models and multilingual training, and CLKD clearly rivals and regularly outperforms the supervised finetuning with gold-labeled target language data. These results are especially surprising for a method that does not require any manual annotation as TyDi-AS2 has diverse questions and documents in native languages and orders of magnitude more data than Xtr-WikiQA does.

Unlike translationese, however, the English teacher's performance is no longer an upper bound as the student models are trained on original language data. In fact, the XLM-R-Large model with both the supervised finetuning and CLKD using data from all the languages (ALL) consistently outperforms the MT+TEACHER pipeline for all the considered target languages. The improved results in Table 6 v/s Table 5 also confirm the importance of training on original target language data as opposed to translationese.

Since the cost of manual annotation makes supervised finetuning infeasible for AS2 tasks, these results demonstrate the advantages of our proposed approach in being able to leverage strong English AS2 models and original target language data. While supervision from a strong English teacher precludes the need for costly manual annotation, when used with original target language data, the teacher absorbs the errors and artifacts introduced by machine translation allowing the student to be

Language	Student LM	Method	bn	fi	id	ja	ko	ru	sw
	MT + TEACHER		63.9	69.2	81.0	55.4	77.8	66.8	86.4
SINGLE	Monolingual	FINETUNE	54.9	61.8	71.2	30.1	71.6	65.0	71.3
		CLKD[E]	53.1	63.0	72.8	34.0	77.7	63.6	74.5
	mBERT	FINETUNE	50.5	60.3	75.5	53.2	70.9	60.1	73.8
		CLKD[E]	53.6	61.6	78.6	57.0	75.9	58.5	78.1
	XLM-R-Base	FINETUNE	50.3	53.8	71.0	45.4	69.5	54.9	69.9
		CLKD[E]	51.5	56.8	77.6	49.9	74.2	54.0	77.5
	XLM-R-Large	FINETUNE	66.4	68.7	78.3	59.3	75.4	65.6	84.3
		CLKD[E]	67.7	67.0	81.9	57.7	80.3	66.4	86.1
ALL	mBERT	FINETUNE	58.5	65.1	80.4	56.6	76.7	63.4	80.6
		CLKD[E]	59.2	63.4	81.8	57.1	81.9	62.6	82.3
	XLM-R-Base	FINETUNE	55.4	60.7	77.1	50.1	73.2	59.2	80.4
		CLKD[E]	56.9	59.8	78.0	52.7	78.0	59.7	83.1
	XLM-R-Large	FINETUNE	**70.0**	**72.3**	82.2	**62.9**	80.5	**68.4**	**88.6**
		CLKD[E]	68.0	68.8	**84.0**	58.0	**83.4**	68.3	87.1

Table 6: **TyDi-AS2**: Averaged test results (P@1) of models trained on original language data for 1) single target language and 2) all the target languages. We highlight better results (FINETUNE vs. CLKD) and additionally use a bold font for **the best student model** for each of the seven target languages.

trained directly on native text. Moreover, the soft-label supervision from the teacher seems even more useful than gold labels for mono-lingual training and/or smaller student models.

7 Conclusion

In this work, we proposed cross-lingual knowledge distillation (CLKD) to leverage strong English AS2 models to train accurate models for low-resource languages without the need for costly manual annotation. Furthermore, we introduced 1) Xtr-WikiQA, a machine-translated WikiQA dataset in 9 additional languages and 2) TyDi-AS2, a new multilingual AS2 benchmark spanning 8 languages.

We conducted comprehensive experiments involving various teachers, students, and training settings, to discuss the potential of CLKD. Our results demonstrate the benefits of using soft supervision from a strong English teacher to train a student model for low-resource languages, suggesting the importance of original target language data compared to translationese potentially due to cultural biases and noise introduced by machine translation.

Despite requiring no manual annotations, CLKD leverages both strong English teachers and original target language data and outperforms or rivals strong baselines such as supervised finetuning with the same amount of data and direct usage of a strong teacher model on English translations.

The results also suggest that CLKD has a potential to greatly reduce the cost of training strong AS2 models for languages lacking labeled training data. To engage studies on AS2 for such languages, we publish Xtr-WikiQA [1] and TyDi-AS2.[2]

Limitations

The proposed CLKD is technically applicable to other NLP tasks, but we discuss the effectiveness of the approach for question answering systems, specifically for answer sentence selection (AS2) tasks. In this study, we put our focus on AS2 tasks as the research community has not well discussed or proposed multilingual AS2 tasks/datasets. We also find that using only English teacher models is another major limitation of this study. However, choices of teacher models in the proposed CLKD are not limited to English models. It would be interesting to discuss the generalizability of the proposed CLKD beyond AS2 tasks, but we note that such discussions would need a much more space to conduct as comprehensive experiments for different NLP tasks as we did for AS2 tasks.

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pages 547–564, Online. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Daniele Bonadiman and Alessandro Moschitti. 2020. A Study on Efficiency, Accuracy and Document Structure for Answer Sentence Selection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5211–5222.

Ruey-Cheng Chen, Evi Yulianti, Mark Sanderson, and W Bruce Croft. 2017. On the Benefit of Incorporating External Features in a Neural Architecture for Answer Sentence Selection. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1017–1020.

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

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020b. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *ArXiv preprint*, abs/1912.09582.

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.

Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. 2022. Pre-training transformer models with sentence-level objectives for answer sentence selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11806–11816.

Matteo Gabburo, Rik Koncel-Kedziorski, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. 2022. Knowledge Transfer from Answer Ranking to Answer Generation. *ArXiv preprint*, abs/2210.12865.

Siddhant Garg and Alessandro Moschitti. 2021. Will this question be answered? question filtering via answer model distillation for efficient question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7329–7346, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. TANDA: transfer and adapt pre-trained transformer models for answer sentence selection. 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 7780–7788. AAAI Press.

Rujun Han, Luca Soldaini, and Alessandro Moschitti. 2021. Modeling context in answer sentence selection systems on a latency budget. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3005–3010, Online. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *ArXiv preprint*, abs/1503.02531.

Chao-Chun Hsu, Eric Lind, Luca Soldaini, and Alessandro Moschitti. 2021. Answer generation for retrieval-based question answering systems. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4276–4282, Online. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2020. Cross-lingual text classification with minimal resources by transferring a sparse teacher. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3604–3622, Online. Association for Computational Linguistics.

Vivek Krishnamurthy, Thuy Vu, and Alessandro Moschitti. 2021. Reference-based Weak Supervision for Answer Sentence Selection using Web Data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4294–4299.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *ArXiv preprint*, abs/1905.07213.

Md Tahmid Rahman Laskar, Xiangji Huang, and Enamul Hoque. 2020. Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5505–5514.

Ivano Lauriola and Alessandro Moschitti. 2021. Answer Sentence Selection Using Local and Global Context in Transformer Models. In *European Conference on Information Retrieval*, pages 298–312. Springer.

Yulong Li, Martin Franz, Md Arafat Sultan, Bhavani Iyer, Young-Suk Lee, and Avirup Sil. 2022. Learning cross-lingual IR from an English retriever. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4428–4436, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv preprint*, abs/1907.11692.

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.

Wenhao Lu, Jian Jiao, and Ruofei Zhang. 2020. Twinbert: Distilling knowledge to twin-structured compressed BERT models for large-scale retrieval. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2645–2652. ACM.

Yoshitomo Matsubara, Luca Soldaini, Eric Lind, and Alessandro Moschitti. 2022. Ensemble Transformer for Efficient and Accurate Ranking Tasks: an Application to Question Answering Systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7259–7272.

Yoshitomo Matsubara, Thuy Vu, and Alessandro Moschitti. 2020. Reranking for efficient transformer-based answer selection. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1577–1580. ACM.

Khalil Mrini, Emilia Farcas, and Ndapandula Nakashole. 2021. Recursive tree-structured self-attention for answer sentence selection. 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 4651–4661.

Massimo Nicosia and Alessandro Moschitti. 2018. Semantic Linking in Convolutional Neural Networks for Answer Sentence Selection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1076.

Geondo Park, Gyeongman Kim, and Eunho Yang. 2021a. Distilling linguistic context for language model compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 364–378, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, et al. 2021b. KLUE: Korean Language Understanding Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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, Zachary 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. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

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.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Luca Soldaini and Alessandro Moschitti. 2020. The cascade transformer: an application for efficient answer sentence selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5697–5708, Online. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Chuanqi Tan, Furu Wei, Qingyu Zhou, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2017. Context-Aware Answer Sentence Selection With Hierarchical Gated Recurrent Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):540–549.

Zhixing Tian, Yuanzhe Zhang, Xinwei Feng, Wenbin Jiang, Yajuan Lyu, Kang Liu, and Jun Zhao. 2020. Capturing Sentence Relations for Answer Sentence Selection with Multi-Perspective Graph Encoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9032–9039.

Kateryna Tymoshenko and Alessandro Moschitti. 2018. Cross-pair text representations for answer sentence selection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2162–2173, Brussels, Belgium. 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.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *ArXiv preprint*, abs/1912.07076.

Thuy Vu and Alessandro Moschitti. 2021. Multilingual Answer Sentence Reranking via Automatically Translated Data. *ArXiv preprint*, abs/2102.10250.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, Suntec, Singapore. Association for Computational Linguistics.

Di Wang and Eric Nyberg. 2015. A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 707–712.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. 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 843–857, Suzhou, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Vancouver, Canada. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zeyu Zhang, Thuy Vu, and Alessandro Moschitti. 2021. Joint Models for Answer Verification in Question Answering Systems. 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 3252–3262.

Language	Student LM	Method	ar	de	es	fr	hi	it	ja	nl	pt
SINGLE	Monolingual	CLKD[E]	7	5	N/A	N/A	1	5	5	7	5
		CLKD[R]	7	7	N/A	N/A	1	3	5	3	3
	mBERT	CLKD[E]	5	3	7	5	5	5	7	3	7
		CLKD[R]	3	7	7	3	5	5	3	7	1
	XLM-R-Base	CLKD[E]	7	7	5	5	5	5	7	5	5
		CLKD[R]	7	7	5	3	5	5	5	1	3
	XLM-R-Large	CLKD[E]	3	5	3	7	3	7	5	3	5
		CLKD[R]	7	7	5	5	3	7	3	3	7
ALL	mBERT	CLKD[E]	3	5	3	3	3	3	3	5	5
		CLKD[R]	7	5	3	3	3	3	5	5	7
	XLM-R-Base	CLKD[E]	3	5	7	1	5	5	3	3	5
		CLKD[R]	7	5	5	5	3	3	3	5	7
	XLM-R-Large	CLKD[E]	7	7	7	3	1	5	5	5	7
		CLKD[R]	3	7	3	1	5	5	7	7	5

Table 7: **Xtr-WikiQA**: Best temperature τ for CLKD we found in search space (see § 5.4) with respect to dev split and used to report test results in Table 4.

A Dataset Validation

Since sentence tokenization and identifying answer sentences can introduce errors, we conducted a manual validation of the TyDi-AS2 datasets. For each language, we randomly selected 50 instances and verified the accuracy of the answer sentences through manual inspection. Our findings revealed that the answer sentences were accurate in 98% of the cases.

B Temperatures for CLKD

Tables 7 - 9 present the best hyperparameter value of temperature τ in CLKD for each configuration for Xtr-WikiQA, Xtr-TyDi-AS2, and TyDi-AS2 datsaets. Following (Matsubara et al., 2022), we select the best temperature value in terms of mean average precision for the development split. Those hyperparameter values are used to obtain student models presented in Tables 4 - 6.

Language	Student LM	Method	bn	fi	id	ja	ko	ru	sw
SINGLE	Monolingual	CLKD[E]	1	3	3	3	3	3	3
	mBERT	CLKD[E]	1	3	3	5	3	3	3
	XLM-R-Base	CLKD[E]	3	3	5	3	3	3	1
	XLM-R-Large	CLKD[E]	3	5	5	3	1	5	3
ALL	mBERT	CLKD[E]	3	3	3	3	3	3	1
	XLM-R-Base	CLKD[E]	3	3	3	3	5	3	1
	XLM-R-Large	CLKD[E]	5	7	7	7	1	5	7

Table 8: **Xtr-TyDi-AS2 (translationese)**: Best temperature τ for CLKD we found in search space (see § 5.4) with respect to dev split and used to report test results in Table 5.

Language	Student LM	Method	bn	fi	id	ja	ko	ru	sw
SINGLE	Monolingual	CLKD[E]	1	3	1	3	5	3	3
	mBERT	CLKD[E]	1	3	3	3	1	3	3
	XLM-R-Base	CLKD[E]	3	3	3	3	3	3	3
	XLM-R-Large	CLKD[E]	1	1	3	1	3	3	3
ALL	mBERT	CLKD[E]	1	3	1	3	1	3	3
	XLM-R-Base	CLKD[E]	3	1	3	3	1	3	1
	XLM-R-Large	CLKD[E]	1	3	3	5	1	3	1

Table 9: **TyDi-AS2 (original)**: Best temperature τ for CLKD we found in search space (see § 5.4) with respect to dev split and used to report test results in Table 6.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We described the limitations of our work in "Limitation" section before "References".

☐ A2. Did you discuss any potential risks of your work?
Not applicable. In this study, we did not introduce new risks besides those of pretrained language models and AS2 datasets we used.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
We summarized the paper's main claims in the abstract and "Introduction" section.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We described datasets and models in Sections 4 and 5.1, respectively.

☑ B1. Did you cite the creators of artifacts you used?
We cited the papers, software, and/or repositories of datasets and models in Sections 4, 5.1, and 5.4, and Table 3, respectively.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. We converted the existing datasets into multilingual AS2 datasets and did not add new contexts to the new datasets except machine-translated texts.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 4.1 and 4.2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Tables 1 and 2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
For the number of model parameters, see Table 3. We provide the computing infrastructure in Section 5.4. We did not track the total computational budget.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 5.4, we explained hyperparameters and search space. Appendix B provides the best-found hyperparameter value from the search space for each of the configurations.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Table 3, Sections 4.1, 4.2 and 5.4.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.