# Cross-Lingual Text Classification with Minimal Resources by Transferring a Sparse Teacher

**Giannis Karamanolakis, Daniel Hsu, Luis Gravano**
Columbia University, New York, NY 10027, USA
{gkaraman, djhsu, gravano}@cs.columbia.edu

## Abstract

Cross-lingual text classification alleviates the need for manually labeled documents in a target language by leveraging labeled documents from other languages. Existing approaches for transferring supervision across languages require *expensive* cross-lingual resources, such as parallel corpora, while less expensive cross-lingual representation learning approaches train classifiers *without* target labeled documents. In this work, we propose a cross-lingual teacher-student method, CLTS, that generates "weak" supervision in the target language using *minimal* cross-lingual resources, in the form of a small number of word translations. Given a limited translation budget, CLTS extracts and transfers only the most important task-specific seed words across languages and initializes a teacher classifier based on the translated seed words. Then, CLTS iteratively trains a more powerful student that also exploits the context of the seed words in *unlabeled* target documents and outperforms the teacher. CLTS is simple and surprisingly effective in 18 diverse languages: by transferring just 20 seed words, even a bag-of-words logistic regression student outperforms state-of-the-art cross-lingual methods (e.g., based on multilingual BERT). Moreover, CLTS can accommodate any type of student classifier: leveraging a *monolingual* BERT student leads to further improvements and outperforms even more expensive approaches by up to 12% in accuracy. Finally, CLTS addresses emerging tasks in low-resource languages using just a small number of word translations.

## 1 Introduction

The main bottleneck in using supervised learning for multilingual document classification is the high cost of obtaining labeled documents for all of the target languages. To address this issue in a target language $L_T$, we consider a cross-lingual text classification approach that requires labeled documents only in a source language $L_S$ and not in $L_T$.

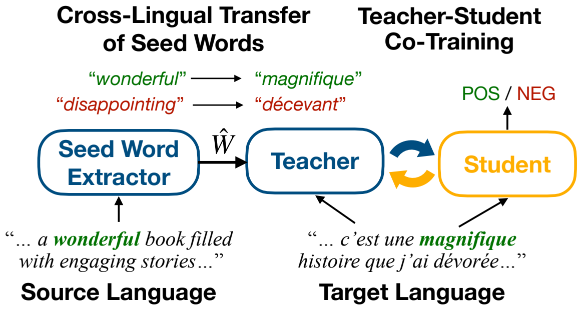

Cross-Lingual Transfer of Seed Words
Teacher-Student Co-Training
"wonderful" → "magnifique"
"disappointing" → "décevant"
POS / NEG
Seed Word Extractor
$\hat{W}$
Teacher
Student
"... a **wonderful** book filled with engaging stories..."
"... c'est une **magnifique** histoire que j'ai dévorée..."
Source Language
Target Language

Figure 1: Our cross-lingual teacher-student (CLTS) method trains a student classifier in the target language by transferring weak supervision across languages.

Existing approaches for transferring supervision across languages rely on large parallel corpora or machine translation systems, which are expensive to obtain and are not available for many languages.[1] To scale beyond high-resource languages, multilingual systems have to reduce the cross-lingual requirements and operate under a limited budget of cross-lingual resources. Such systems typically ignore target-language supervision, and rely on feature representations that bridge languages, such as cross-lingual word embeddings (Ruder et al., 2019) or multilingual transformer models (Wu and Dredze, 2019; Pires et al., 2019). This general approach is less expensive but has a key limitation: by not considering labeled documents in $L_T$, it may fail to capture predictive patterns that are specific to $L_T$. Its performance is thus sensitive to the quality of pre-aligned features (Glavaš et al., 2019).

In this work, we show how to obtain weak supervision for training accurate classifiers in $L_T$ without using manually labeled documents in $L_T$

[1] Google Translate (https://translate.google.com/) is available for 103 out of the about 4,000 written languages (https://www.ethnologue.com/).

*Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3604–3622
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

or expensive document translations. We propose a novel approach for cross-lingual text classification that transfers weak supervision from $L_S$ to $L_T$ using *minimal* cross-lingual resources: we only require a small number of task-specific keywords, or seed words, to be translated from $L_S$ to $L_T$. Our core idea is that the most indicative seed words in $L_S$ often translate to words that are also indicative in $L_T$. For instance, the word "wonderful" in English indicates positive sentiment, and so does its translation "magnifique" in French. Thus, given a limited budget for word translations (e.g., from a bilingual speaker), only the most important seed words should be prioritized to transfer task-specific information from $L_S$ to $L_T$.

Having access only to limited cross-lingual resources creates important challenges, which we address with a novel cross-lingual teacher-student method, CLTS (see Figure 1).

**Efficient transfer of supervision across languages:** As a first contribution, we present a method for cross-lingual transfer in low-resource settings with a limited word translation budget. CLTS extracts the most important seed words using the translation budget as a sparsity-inducing regularizer when training a classifier in $L_S$. Then, it transfers seed words and the classifier's weights across languages, and initializes a teacher classifier in $L_T$ that uses the translated seed words.

**Effective training of classifiers without using any labeled target documents:** The teacher, as described above, predicts meaningful probabilities only for documents that contain translated seed words. Because translations can induce errors and the translation budget is limited, the translated seed words may be noisy and not comprehensive for the task at hand. As a second contribution, we extend the "weakly-supervised co-training" method of Karamanolakis et al. (2019) to our cross-lingual setting for training a stronger student classifier using the teacher and unlabeled-only target documents. The student outperforms the teacher across all languages by 59.6%.

**Robust performance across languages and tasks:** As a third contribution, we empirically show the benefits of generating weak supervision in 18 diverse languages and 4 document classification tasks. With as few as 20 seed-word translations and a bag-of-words logistic regression student, CLTS outperforms state-of-the-art methods relying on more complex multilingual models, such as multilingual BERT, across most languages. Using a monolingual BERT student leads to further improvements and outperforms even more expensive approaches (Figure 2). CLTS does not require cross-lingual resources such as parallel corpora, machine translation systems, or pre-trained multilingual language models, which makes it applicable in low-resource settings. As a preliminary exploration, we address medical emergency situation detection in Uyghur and Sinhalese with just 50 translated seed words per language, which could be easily obtained from bilingual speakers.

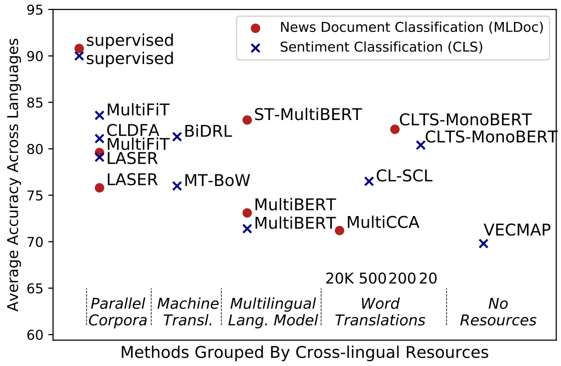


Figure 2: CLTS leverages a small number of word translations more effectively than previous methods and sometimes outperforms more expensive methods. (Refer to Sections 2 and 5 for details.)

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 defines our problem of focus. Section 4 presents our cross-lingual teacher-student approach.[2] Section 5 describes our experimental setup and results. Finally, Section 6 concludes and suggests future work.

## 2 Related Work

Relevant work spans cross-lingual text classification and knowledge distillation.

### 2.1 Cross-Lingual Text Classification

We focus on a cross-lingual text classification scenario with labeled data in the source language $L_S$ and unlabeled data in the target language $L_T$. We review the different types of required cross-lingual resources, starting with the most expensive types.

**Annotation Projection and Machine Translation.** With parallel corpora (i.e., corpora where

[2] Our Python implementation is publicly available at https://github.com/gkaramanolakis/clts.

each document is written in both $L_S$ and $L_T$), a classifier trained in $L_S$ predicts labels for documents in $L_S$ and its predictions are projected to documents in $L_T$ to train a classifier in $L_T$ (Mihalcea et al., 2007; Rasooli et al., 2018). Unfortunately, parallel corpora are hard to find, especially in low-resource domains and languages.

Without parallel corpora, documents can be translated using machine translation (MT) systems (Wan, 2008, 2009; Salameh et al., 2015; Mohammad et al., 2016). However, high-quality MT systems are limited to high-resource languages. Even when an MT system is available, translations may change document semantics and degrade classification accuracy (Salameh et al., 2015).

**Cross-Lingual Representation Learning.** Other approaches rely on less expensive resources to align feature representations across languages, typically in a shared feature space to enable cross-lingual model transfer.

Cross-lingual word embeddings, or CLWE, represent words from different languages in a joint embedding space, where words with similar meanings obtain similar vectors regardless of their language. (See Ruder et al. (2019) for a survey.) Early CLWE approaches required expensive parallel data (Klementiev et al., 2012; Täckström et al., 2012). In contrast, later approaches rely on high-coverage bilingual dictionaries (Gliozzo and Strapparava, 2006; Faruqui and Dyer, 2014; Gouws et al., 2015) or smaller "seed" dictionaries (Gouws and Søgaard, 2015; Artetxe et al., 2017). Some recent CLWE approaches require no cross-lingual resources (Lample et al., 2018; Artetxe et al., 2018; Chen and Cardie, 2018; Søgaard et al., 2018) but perform substantially worse than approaches using seed dictionaries of 500-1,000 pairs (Vulić et al., 2019). Our approach does not require CLWE and achieves competitive classification performance with substantially fewer translations of *task-specific* words.

Recently, multilingual transformer models were pre-trained in multiple languages in parallel using language modeling objectives (Devlin et al., 2019; Conneau and Lample, 2019). Multilingual BERT, a version of BERT (Devlin et al., 2019) that was trained on 104 languages in parallel without using any cross-lingual resources, has received significant attention (Karthikeyan et al., 2019; Singh et al., 2019; Rogers et al., 2020). Multilingual BERT performs well on zero-shot cross-lingual transfer (Wu and Dredze, 2019; Pires et al., 2019) and its performance can be further improved by considering target-language documents through self-training (Dong and de Melo, 2019). In contrast, our approach does not require multilingual language models and sometimes outperforms multilingual BERT using a *monolingual* BERT student.

### 2.2 Knowledge Distillation

Our teacher-student approach is related to "knowledge distillation" (Buciluă et al., 2006; Ba and Caruana, 2014; Hinton et al., 2015), where a student classifier is trained using the predictions of a teacher classifier. Xu and Yang (2017) apply knowledge distillation for cross-lingual text classification but require expensive parallel corpora. MultiFiT (Eisenschlos et al., 2019) trains a classifier in $L_T$ using the predictions of a cross-lingual model, namely, LASER (Artetxe and Schwenk, 2019), that also requires large parallel corpora. Vyas and Carpuat (2019) classify the semantic relation (e.g., synonymy) between two words from different languages by transferring *all* training examples across languages. Our approach addresses a different problem, where training examples are full documents (not words), and transferring source training documents would require MT. Related to distillation is the semi-supervised approach of Shi et al. (2010) that trains a target classifier by transferring a source classifier using high-coverage dictionaries. Our approach is similar, but trains a classifier using sparsity regularization, and translates only the most important seed words.

## 3 Problem Definition

Consider a source language $L_S$, a target language $L_T$, and a classification task with $K$ predefined classes of interest $\mathcal{Y} = \{1, \ldots, K\}$ (e.g., sentiment categories). Labeled documents $D_S = \{(x_i^S, y_i)\}_{i=1}^N$ are available in $L_S$, where $y_i \in \mathcal{Y}$ and each source document $x_i^S$ is a sequence of words from the source vocabulary $V_S$. Only unlabeled documents $D_T = \{x_i^T\}_{i=1}^M$ are available in $L_T$, where each target document $x_i^T$ is a sequence of words from the target vocabulary $V_T$. We assume that there is no significant shift in the conditional distribution of labels given documents across languages. Furthermore, we assume a limited translation budget, so that up to $B$ words can be translated from $L_S$ to $L_T$.

Our goal is to use the labeled source documents $D_S$, the unlabeled target documents $D_T$, and the translations of no more than $B$ source words to train a classifier that, given an unseen test document $x_i^T$ in the target language $L_T$, predicts the corresponding label $y_i \in \mathcal{Y}$.

## 4 Cross-Lingual Teacher-Student

We now describe our cross-lingual teacher-student method, CLTS, for cross-lingual text classification. Given a limited budget of $B$ translations, CLTS extracts only the $B$ most important seed words in $L_S$ (Section 4.1). Then, CLTS transfers the seed words and their weights from $L_S$ to $L_T$, to initialize a classifier in $L_T$ (Section 4.2). Using this classifier as a teacher, CLTS trains a student that predicts labels using both seed words and their context in target documents (Section 4.3).

### 4.1 Seed-Word Extraction in $L_S$

CLTS starts by automatically extracting a set $G_k^S$ of indicative seed words per class $k$ in $L_S$. Previous extraction approaches, such as tf-idf variants (Angelidis and Lapata, 2018), have been effective in monolingual settings with limited labeled data. In our scenario, with sufficiently many labeled *source* documents and a limited translation budget $B$, we propose a different approach based on a supervised classifier trained with sparsity regularization.

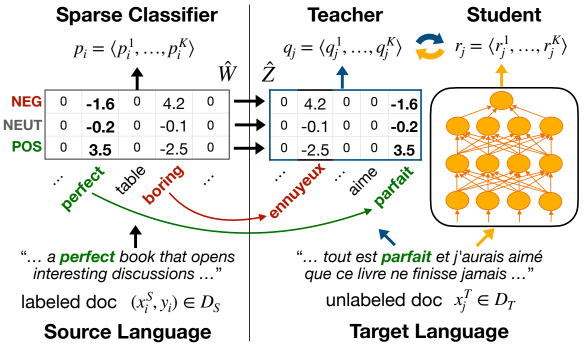
Specifically, CLTS extracts seed words from the weights $W \in \mathbb{R}^{K \times |V_S|}$ of a classifier trained using $D_S$. Given a source document $x_i^S$ with a bag-of-words encoding $h_i^S \in \mathbb{R}^{|V_S|}$, the classifier predicts class probabilities $p_i = \langle p_i^1, \ldots, p_i^K \rangle = \text{softmax}(Wh_i)$. CLTS includes the word $v_c \in V_S$ in $G_k^S$ if the classifier considers it to increase the probability $p_i^k$ through a positive weight $W_{kc}$:

$$G_k^S = \{v_c^S \mid W_{kc} > 0\}. \tag{1}$$

The set of all source seed words $G^S = G_1^S \cup \cdots \cup G_K^S$ may be much larger than the translation budget $B$. We encourage the classifier to capture only the most important seed words *during* training through sparsity regularization:

$$\hat{W} = \underset{W}{\arg\min} \sum_{i=1}^{N} \mathcal{L}(y_i, Wh_i^S) + \lambda_B \mathcal{R}_{\text{sparse}}(W) \tag{2}$$

where $\mathcal{L}$ is the training loss function (logistic loss), $\mathcal{R}_{\text{sparse}}(.)$ is a sparsity regularizer (L1 norm), and $\lambda_B \in \mathbb{R}$ is a hyperparameter controlling the relative power of $\mathcal{R}_{\text{sparse}}$. Higher $\lambda_B$ values lead to sparser matrices $\hat{W}$ and thus to fewer seed words.



Sparse Classifier
Teacher
Student
$p_i = \langle p_i^1, \ldots, p_i^K \rangle$
$q_j = \langle q_j^1, \ldots, q_j^K \rangle$
$r_j = \langle r_j^1, \ldots, r_j^K \rangle$
$\hat{W}$
$\hat{Z}$
NEG
NEUT
POS
0 -1.6 0 4.2 0
0 -0.2 0 -0.1 0
0 3.5 0 -2.5 0
0 4.2 0 0 -1.6
0 -0.1 0 0 -0.2
0 -2.5 0 0 3.5
… perfect table boring …
… ennuyeux … aime parfait
"… a perfect book that opens interesting discussions …"
labeled doc $(x_i^S, y_i) \in D_S$
Source Language
"… tout est parfait et j'aurais aimé que ce livre ne finisse jamais …"
unlabeled doc $x_j^T \in D_T$
Target Language

Figure 3: CLTS (1) learns a sparse weight matrix $\hat{W}$ in $L_S$; (2) transfers the columns of $\hat{W}$ for $B$ seed words to initialize $\hat{Z}$; and (3) uses $\hat{Z}$ as a teacher to iteratively train a student on unlabeled documents $D_T$.

Therefore, we tune[3] $\lambda_B$ to be as high as possible while at the same time leading to the extraction of at least $B$ seed words. After training, $G^S$ consists of the $B$ seed words with highest weight.

### 4.2 Cross-Lingual Seed Weight Transfer

We now describe our cross-lingual transfer method. CLTS transfers both translated seed words and their learned weights to initialize a "weak" classifier in $L_T$ that considers translated seed words and their relative importance for the target task.

Specifically, CLTS first translates the $B$ seed words in $G^S$ into a set $G^T$ with seed words in $L_T$. Then, for each translation pair $(v^S, v^T)$, CLTS transfers the column for $v^S$ in $\hat{W}$ to a corresponding column for $v^T$ in a $K \times |V_T|$ matrix $\hat{Z}$:

$$\hat{Z}_{k,v^S} = \hat{W}_{k,v^T} \quad \forall k \in [K] \tag{3}$$

Importantly, for each word, we transfer the weights for all classes (instead of just a single weight $\hat{W}_{kc}$) across languages. Therefore, *without using any labeled documents* in $L_T$, CLTS constructs a classifier that, given a test document $x_j^T$ in $L_T$, predicts class probabilities $q_j = \langle q_j^1, \ldots, q_j^K \rangle$:

$$q_j^k = \frac{\exp\left(\hat{z}_k^\top h_j^T\right)}{\sum_{k'} \exp\left(\hat{z}_{k'}^\top h_j^T\right)}, \tag{4}$$

where $h_j^T \in \mathbb{R}^{|V_T|}$ is a bag-of-words encoding for $x_j^T$ and $\hat{z}_k$ is the $k$-th row of $\hat{Z}$. Note that columns of $\hat{Z}$ for non-seed words in $V_T$ are all zeros and thus this classifier predicts meaningful probabilities only for documents with seed words in $G^T$.

[3] We efficiently tune $\lambda_B$ by computing the "regularization path" with the "warm-start" technique (Koh et al., 2007).

### 4.3 Teacher-Student Co-Training in $L_T$

We now describe how CLTS trains a classifier in $L_T$ that leverages indicative features, which may not be captured by the small set of translated seed words. As illustrated in Figure 3, translated seed words (e.g., "parfait") often co-occur with other words (e.g., "aime," meaning "love") that have zero weight in $\hat{Z}$ but are also helpful for the task at hand. To exploit such words in the absence of labeled target documents, we extend the monolingual weakly-supervised co-training method of Karamanolakis et al. (2019) to our cross-lingual setting, and use our classifier based on translated seed words as a teacher to train a student, as we describe next.

First, CLTS uses our classifier from Equation 5 as a teacher to predict labels $q_j$ for *unlabeled* documents $x_j^T \in D_T$ that contain seed words: $D'_T = \{(x_j^T, q_j)\}_{x_j^T | x_j^T \cap G^T \neq \varnothing} \subseteq D_T$. Note that our teacher with weights transferred across languages is different than that of Karamanolakis et al. (2019), which simply "counts" seed words.

Next, CLTS trains a student $f^T$ that also exploits the context of the seed words. Given a document $x_j^T$ in $L_T$, the student predicts class probabilities:

$$r_j = \langle r_j^1, \ldots, r_j^K \rangle = f^T(x_j^T; \theta), \qquad (5)$$

where the predictor function $f^T$ with weight parameters $\theta$ can be of any type, such as a pre-trained transformer-based classifier that captures language-specific word composition. The student is trained via the "distillation" objective:

$$\hat{\theta} = \arg\min_{\theta} \sum_{(x_j^T, q_j) \in D'_T} H(q_j, f^T(x_j^T)) + \lambda \mathcal{R}(\theta), \qquad (6)$$

where $H(q, r) = -\sum_k q^k \log r^k$ is the cross entropy between student's and teacher's predictions, $\mathcal{R}(.)$ is a regularizer (L2 norm), and $\lambda \in \mathbb{R}$ is a hyperparameter controlling the relative power of $\mathcal{R}$. Importantly, through extra regularization ($\mathcal{R}$, dropout) the student also associates non-seed words with target classes, and generalizes better than the teacher by making predictions even for documents that do not contain any seed words.

Then, CLTS uses the student in place of the teacher to annotate *all* $M$ unlabeled examples in $D_T$ and create $D'_T = \{(x_j^T, \hat{f}^T(x_j^T))\}_{j \in [M]}$. While in the first iteration $D'_T$ contains only documents with seed words, in the second iteration CLTS adds in $D'_T$ *all* unlabeled documents to create a larger training set for the student. This also differs from Karamanolakis et al. (2019), which updates the weights of the initial seed words but does not provide pseudo-labels for documents with no seed words. This change is important in our cross-lingual setting with a limited translation budget, where the translated seed words $G^T$ may only cover a very small subset $D'_T$ of $D_T$.

**Algorithm 1** Cross-Lingual Teacher-Student

```
Input: Unlabeled documents $D_T = \{x_j^T\}_{j=1}^M$, labeled documents $D_S = \{(x_i^S, y_i)\}_{i=1}^N$, budget of up to $B$ word translations ($L_S$ to $L_T$)
Output: $\hat{f}^T$: predictor function in $L_T$
Learn $\lambda_B$-sparse $\hat{W}$ using $D_S$, $B$ (Eq. (2))
Extract $B$ seed words $G^S$ from $\hat{W}$ (Eq. (1))
Translate $G^S$ to target seed words $G^T$ in $L_T$
Transfer $\hat{W}$ to initialize teacher $\hat{Z}$ (Eq. (3))
Get $D'_T = \{(x_j^T, q_j)\}_{x_j^T | x_j^T \cap G^T \neq \varnothing}$ (Eq. (4))
Repeat until convergence
    a. Learn student $\hat{f}^T$ using $D'_T$ (Eq. (6))
    b. Get $D'_T = \{(x_j^T, \hat{f}^T(x_j^T))\}_{j \in [M]}$ (Eq. (5))
```

Algorithm 1 summarizes the CLTS method for cross-lingual classification by translating $B$ seed words. Iterative co-training converges when the disagreement between the student's and teacher's hard predictions on unlabeled data stops decreasing. In our experiments, just two rounds of co-training are generally sufficient for the student to outperform the teacher and achieve competitive performance even with a tight translation budget $B$.

## 5 Experiments

We now evaluate CLTS for several cross-lingual text classification tasks in various languages.

### 5.1 Experimental Settings

**Datasets:** We use English (En) as a source language, and evaluate CLTS on 18 diverse target languages: Bulgarian (Bg), German (De), Spanish (Es), Persian (Fa), French (Fr), Croatian (Hr), Hungarian (Hu), Italian (It), Japanese (Ja), Polish (Pl), Portuguese (Pt), Russian (Ru), Sinhalese (Si), Slovak (Sk), Slovenian (Sl), Swedish (Sv), Uyghur (Ug), and Chinese (Zh). We focus on four classification tasks: **T1:** 4-class classification of news documents in the MLDoc corpus (Schwenk and Li, 2018); **T2:** binary sentiment classification of product reviews in the CLS corpus (Prettenhofer and Stein, 2010); **T3:** 3-class sentiment classification of tweets in the Twitter Sentiment corpus (TwitterSent; Mozetič et al. (2016)), Persian reviews in the SentiPers corpus (Hosseini et al., 2018), and Uyghur documents in the LDC LORELEI corpus (Strassel and Tracey, 2016); and **T4:** medical emergency situation detection in Uyghur and Sinhalese documents from the LDC LORELEI corpus. The appendix discusses additional dataset details.

**Experimental Procedure:** We use English as the source language, where we train a source classifier and extract $B$ seed words using labeled documents (Section 4.1). Then, we obtain translations for $B \leq 500$ English seed words using the MUSE[4] bilingual dictionaries (Lample et al., 2018). For Uyghur and Sinhalese, which have no entries in MUSE, we translate seed words through Google Translate.[5] The appendix reports additional seed-word translation details. We do not use labeled documents in the target language for training (Section 3). We report both the teacher's and student's performance in $L_T$ averaged over 5 different runs. We consider any test document that contains no seed words as a "mistake" for the teacher.

**Model Configuration:** For the student, we experiment with a bag-of-ngrams ($n = 1, 2$) logistic regression classifier (LogReg) and a linear classifier using pre-trained monolingual BERT embeddings as features (MonoBERT; Devlin et al. (2019)). The appendix lists the implementation details. We do not optimize any hyperparameters in the target language, except for $B$, which we vary between 6 and 500 to understand the impact of translation budget on performance. CLS does not contain validation sets, so we fix $B = 20$ and translate 10 words for each of the two sentiment classes. More generally, to cover all classes we extract and translate $\frac{B}{K}$ seed words per class. We perform two rounds of teacher-student co-training, which provided most of the improvement in Karamanolakis et al. (2019).

**Model Comparison:** For a robust evaluation of CLTS, we compare models with different types of cross-lingual resources. ***Project-**** uses the parallel LDC or EuroParl (EP) corpora for annotation projection (Rasooli et al., 2018). ***LASER*** uses millions of parallel corpora to obtain cross-lingual sentence embeddings (Artetxe and Schwenk, 2019). ***MultiFiT*** uses ***LASER*** to create pseudo-labels in $L_T$ (Eisenschlos et al., 2019) and trains a classifier in $L_T$ based on a pre-trained language model (Howard and Ruder, 2018). ***CLWE-par*** uses parallel corpora to train CLWE (Rasooli et al., 2018). ***MT-BOW*** uses Google Translate to translate test documents from $L_T$ to $L_S$ and applies a bag-of-words classifier in $L_S$ (Prettenhofer and Stein, 2010). ***BiDRL*** uses Google Translate to translate documents from $L_S$ to $L_T$ and $L_T$ to $L_S$ (Zhou et al., 2016). ***CLDFA*** uses task-specific parallel corpora for cross-lingual distillation (Xu and Yang, 2017). ***SentiWordNet*** uses bilingual dictionaries with over 20K entries to transfer the SentiWordNet03 (Baccianella et al., 2010) to the target language and applies a rule-based heuristic (Rasooli et al., 2018). ***CLWE-Wikt*** uses bilingual dictionaries with over 20K entries extracted from Wiktionary[6] to create CLWE for training a bi-directional LSTM classifier (Rasooli et al., 2018). ***MultiCCA*** uses bilingual dictionaries with around 20K entries to train CLWE (Ammar et al., 2016), trains a convolutional neural network (CNN) in $L_S$ and applies it in $L_T$ (Schwenk and Li, 2018). ***CL-SCL*** obtains 450 word translations as "pivots" for cross-lingual domain adaptation (Prettenhofer and Stein, 2010). Our CLTS approach uses $B$ word translations not for domain adaptation but to create weak supervision in $L_T$ through the teacher (Teacher) for training the student (Student-LogReg or Student-MonoBERT). ***VECMAP*** uses identical strings across languages as a weak signal to train CLWE (Artetxe et al., 2017). ***MultiBERT*** uses multilingual BERT to train a classifier in $L_S$ and applies it in $L_T$ (Wu and Dredze, 2019) without considering labeled documents in $L_T$ (zero-shot setting). ***ST-MultiBERT*** further considers labeled documents in $L_T$ for fine-tuning multilingual BERT through self-training (Dong and de Melo, 2019). The appendix discusses more comparisons.

### 5.2 Experimental Results

Figure 4 shows results for each classification task and language. The rightmost column of each table reports the average performance across all languages (and domains for CLS). For brevity, we report the average performance across the three review domains (Books, DVD, Music) for each lan-

[4] https://github.com/facebookresearch/MUSE#ground-truth-bilingual-dictionaries

[5] Google Translate started supporting Uyghur on February 26, 2020, and Sinhalese at an earlier (unknown) time.

[6] https://www.wiktionary.org/

| Method | De | Es | Fr | It | Ru | Zh | Ja | AVG |
|---|---|---|---|---|---|---|---|---|
| *Methods below use parallel corpora (MultiFiT requires LASER)* | | | | | | | | |
| LASER | 87.7 | **79.3** | 84.0 | 71.2 | 67.3 | 76.7 | 64.6 | 75.8 |
| MultiFiT | **91.6** | 79.1 | **89.4** | **76.0** | **67.8** | **82.5** | **69.6** | **79.4** |
| *Methods below use pre-trained multi-lingual language models* | | | | | | | | |
| MultiBERT | 79.8 | 72.1 | 73.5 | 63.7 | 73.7 | 76.0 | 72.8 | 73.1 |
| ST-MultiBERT | **90.0** | **85.3** | **88.4** | **75.2** | **79.3** | **87.0** | **76.8** | **83.1** |
| *Methods below use bilingual dictionaries (Student requires Teacher)* | | | | | | | | |
| MultiCCA ($B$=20K) | 81.2 | 72.5 | 72.4 | 69.4 | 60.8 | 74.7 | 67.6 | 71.2 |
| Teacher ($B$=160) | 72.7 | 73.5 | 77.6 | 62.5 | 46.9 | 53.3 | 31.9 | 59.8 |
| Student-LogReg | 87.4 | 86.0 | 89.1 | 70.5 | 71.9 | 82.4 | 68.8 | 79.4 |
| Student-MonoBERT | **90.4** | **86.3** | **91.2** | **74.7** | **75.6** | **84.0** | **72.6** | **82.1** |

(a) Accuracy results on MLDoc.

| Model | De | Fr | Ja | AVG |
|---|---|---|---|---|
| *Methods below use parallel corpora or MT* | | | | |
| MT-BOW | 78.3 | 78.5 | 71.2 | 76.0 |
| BiDRL | 84.3 | 83.5 | 76.2 | 81.3 |
| CLDFA | 82.0 | 83.1 | 78.1 | 81.1 |
| LASER | 80.4 | 82.7 | 75.3 | 79.5 |
| MultiFiT | **85.3** | **85.6** | **79.9** | **83.6** |
| *Methods below use multi-lingual language models* | | | | |
| MultiBERT | 72.0 | 75.4 | 66.9 | 71.4 |
| *Methods below use dictionaries or no resources* | | | | |
| VECMAP | 75.3 | 78.2 | 55.9 | 69.8 |
| CL-SCL ($B$=450) | 78.1 | 78.4 | 73.1 | 76.5 |
| Teacher ($B$=20) | 38.1 | 48.6 | 22.7 | 36.5 |
| Student-LogReg | 78.7 | 79.6 | **78.6** | 79.0 |
| Student-MonoBERT | **80.1** | **83.4** | 77.6 | **80.4** |

(b) Accuracy results on CLS.

| Method | Ar | Bg | De | Es | Fa | Hr | Hu | Pl | Pt | Ru | Sk | Sl | Sv | Ug | AVG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Methods below use parallel corpora* | | | | | | | | | | | | | | | |
| Project-LDC | 37.2 | - | - | **42.7** | 33.1 | - | **47.0** | - | - | **48.0** | - | - | - | **38.6** | (41.1) |
| Project-EP | - | **38.7** | **47.3** | 41.8 | - | - | 38.1 | 38.8 | **39.3** | - | **30.0** | **44.6** | **44.6** | - | (40.4) |
| CLWE-Par | **37.3** | 33.0 | 43.5 | 42.6 | **40.1** | **30.8** | 41.1 | **41.7** | 38.6 | 44.8 | 22.6 | 32.2 | 39.1 | 30.0 | 37.0 |
| *Methods below use comparable corpora or bilingual dictionaries* | | | | | | | | | | | | | | | |
| CLWE-CP | 21.1 | 28.6 | 37.7 | 27.7 | 20.7 | 13.9 | 22.4 | 30.2 | 22.2 | 25.3 | 24.6 | 25.3 | 31.1 | 25.7 | 25.5 |
| SentiWordNet ($B$>20K) | 25.6 | 30.6 | 32.0 | 25.3 | 25.3 | 19.8 | 29.2 | 26.0 | 22.9 | 29.5 | 19.2 | 28.1 | 22.7 | 36.7 | 26.6 |
| CLWE-Wikt ($B$>20K) | 31.0 | 45.3 | 51.0 | 37.7 | 31.7 | - | 40.8 | 32.9 | 35.4 | **43.8** | 36.6 | 32.1 | 40.4 | 28.0 | (37.4) |
| Teacher ($B$=500) | 22.7 | 42.8 | 45.5 | 42.7 | 30.9 | 36.4 | 39.4 | 40.7 | 34.4 | 29.8 | 40.4 | 29.5 | 38.7 | 20.3 | 35.3 |
| Student-LogReg | **39.0** | **46.3** | **52.5** | **44.9** | **45.7** | **39.4** | **45.2** | **45.4** | **38.7** | 43.2 | **43.3** | **42.1** | **50.4** | **41.2** | **44.1** |

(c) Macro-averaged F1 results on TwitterSent, SentiPers, and LORELEI.

Figure 4: Classification results, with methods grouped according to the type of cross-lingual resources required. For some methods, average performance (rightmost column) is in parentheses because it is computed on a subset of languages. Across all datasets, CLTS outperforms other methods that require similar types of cross-lingual resources; in many cases (red) CLTS outperforms even *more expensive* state-of-the-art approaches.

guage in the CLS corpus. The appendix discusses detailed results and ablation experiments.

**Student outperforms Teacher.** Teacher considers the noisy translated seed words for classification. Even the simple Student-LogReg technique leverages the context of the seed words and substantially outperforms Teacher. Leveraging pre-trained representations in Student-MonoBERT leads to further improvement. On average, across all languages and datasets, Student outperforms Teacher by 59.6%: CLTS effectively improves performance in $L_T$ *without using labeled documents*.

**Student outperforms previous approaches.** Student-MonoBERT outperforms *MultiBERT* by 12.5% on average across all languages and domains in MLDoc and CLS: CLTS effectively generates weak supervision in $L_T$ for fine-tuning monolingual BERT. Importantly, CLTS is effective under minimal resources: with the translation of just $\frac{B}{K}$ seed words per class, Student-LogReg outperforms other approaches that rely on much larger dictionaries (*MultiCCA*, *CL-SCL*, *SentiWordNet*, *CLWE-Wiktionary*). Surprisingly, in several languages CLTS outperforms even more expensive approaches that rely on parallel corpora or machine translation systems (*LASER*, *MultiFiT*, *MT-BOW*, *BiDRL*, *CLDFA*, *CLWE-BW*, *Project-LDC*).

**CLTS is effective under a minimal translation budget.** Figure 5 shows CLTS's performance as a function of the number of seed words per class ($\frac{B}{K}$). Even with just 3 seed words per class, Student-MonoBERT performs remarkably well. Student's and Teacher's performance significantly increases with $\frac{B}{K}$ and most performance gains are obtained for lower values of $\frac{B}{K}$. This is explained by the fact that CLTS prioritizes the most indicative seed words for translation. Therefore, as $\frac{B}{K}$ increases, the additional seed words that are translated are less indicative than the already-translated seed words and as a result have lower chances of translating to important seed words in the target language. The gap between the Teacher and Student performance has a maximum value of 40 absolute accuracy points and decreases as Teacher considers more seed words but does not get lower than 10, highlighting that Student learns predictive patterns in

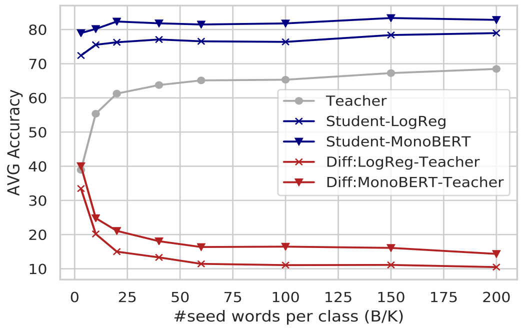


Figure 5: Validation accuracy across all MLDoc languages as a function of the translation budget $\frac{B}{K}$.

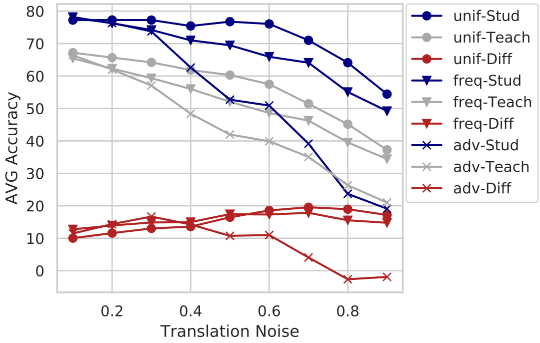
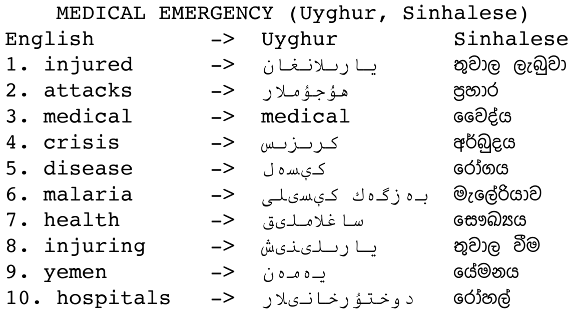


Figure 6: Average validation accuracy in MLDoc for Teacher (Teach), Student-LogReg (Stud), and their absolute difference in accuracy (Diff) under different scales of noise applied to the translated seed words: "unif" replaces a seed word with a different word sampled uniformly at random from $V_T$, "freq" replaces a seed word with a word randomly sampled from $V_T$ with probability proportional to its frequency in $D_T$, "adv" assigns a seed word to a different random class $k' \neq k$ by swapping its class weights in $\hat{Z}$.

$L_T$ that may never be considered by Teacher.

**CLTS is robust to noisy translated seed words.** In practice, an indicative seed word in $L_S$ may not translate to an indicative word in $L_T$. Our results above show that Student in CLTS performs well even when seed words are automatically translated across languages. To further understand our method's behavior with noisy translated seed words, we introduce additional simulated noise of different types and severities. According to Figure 6, "unif" and "freq" noise, which replace translated seed words with random words, affect CLTS less than "adv" noise, which introduces many erroneous teacher-labels. Student is less sensitive than Teacher to noisy seed words: their performance gap (*-Diff) increases with the magnitude of translation noise (up to 0.7) for both "unif" and "freq" noise. Student's accuracy is relatively high for noise rates up to 0.3, even with "adv" noise: CLTS is effective even when 30% of the translated seed words are assumed indicative for the wrong class.



MEDICAL EMERGENCY (Uyghur, Sinhalese)

| English | | Uyghur | Sinhalese |
|---|---|---|---|
| 1. injured | -> | يارىلانغان | තුවාල ලැබුවා |
| 2. attacks | -> | ھۇجۇملار | ප්‍රහාර |
| 3. medical | -> | medical | වෛද්‍ය |
| 4. crisis | -> | كرىزىس | අර්බුදය |
| 5. disease | -> | كېسەل | රෝගය |
| 6. malaria | -> | بەزگەك كېسىلى | මැලේරියාව |
| 7. health | -> | ساغلاملىق | සෞඛ්‍යය |
| 8. injuring | -> | يارىلىنىش | තුවාල වීම |
| 9. yemen | -> | يەمەن | යේමනය |
| 10. hospitals | -> | دوختۇرخانىلار | රෝහල් |

Figure 7: Top 10 extracted seed words for the "medical emergency" class and their translations to Uyghur and Sinhalese. Google Translate erroneously returns "medical" as a Uyghur translation of the word "medical."

### 5.3 Addressing Emerging Classification Tasks in Low-Resource Languages

We now show a preliminary exploration of CLTS for detecting medical emergency situations in the low-resource Uyghur and Sinhalese languages by just translating $B$=50 English seed words across languages. Figure 7 shows the top 10 seed words transferred by CLTS for the medical emergency class. We train Student-LogReg because BERT is not available for Uyghur or Sinhalese. End-to-end training and evaluation of CLTS takes just 160 seconds for Uyghur and 174 seconds for Sinhalese. The accuracy in Uyghur is 23.9% for the teacher and 66.8% for the student. The accuracy in Sinhalese is 30.4% for the teacher and 73.2% for the student. The appendix has more details. These preliminary results indicate that CLTS could be easily applied for emerging tasks in low-resource languages, for example by asking a bilingual speaker to translate a small number of seed words. We expect such correct translations to lead to further improvements over automatic translations.

## 6 Conclusions and Future Work

We presented a cross-lingual text classification method, CLTS, that efficiently transfers weak supervision across languages using minimal cross-lingual resources. CLTS extracts and transfers just a small number of task-specific seed words, and creates a teacher that provides weak supervision for training a more powerful student in the target language. We present extensive experiments on 4 classification tasks and 18 diverse languages, including low-resource languages. Our results show that even a simple student outperforms the teacher and previous state-of-the-art approaches with more complex models and more expensive resources, highlighting the promise of generating weak supervision in the target language. In future work, we plan to extend CLTS for handling cross-domain distribution shift (Ziser and Reichart, 2018) and multiple source languages (Chen et al., 2019). It would also be interesting to combine CLTS with available cross-lingual models, and extend CLTS for more tasks, such as cross-lingual named entity recognition (Xie et al., 2018), by considering teacher architectures beyond bag-of-seed-words.

## Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This material is based upon work supported by the National Science Foundation under Grant No. IIS-15-63785.

## References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Xilun Chen, Ahmed Hassan, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*.

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*.

Xin Dong and Gerard de Melo. 2019. A robust self-learning framework for cross-lingual text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

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*.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2018. Sentipers: A sentiment analysis corpus for Persian. *arXiv preprint arXiv:1801.07737*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*.

Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. 2007. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8:1519–1555.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Twitter sentiment for 15 European languages. Slovenian language resource repository CLARIN.SI.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Mohammad Sadegh Rasooli, Noura Farra, Axinia Radeva, Tao Yu, and Kathleen McKeown. 2018. Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32(1-2):143–165.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language Technologies*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. Bert is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Stephanie Strassel and Jennifer Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Yogarshi Vyas and Marine Carpuat. 2019. Weakly supervised cross-lingual semantic relation classification via knowledge distillation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

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*. Association for Computational Linguistics.

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

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime G Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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*.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive refinement of cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Yftah Ziser and Roi Reichart. 2018. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

# A Appendix

For reproducibility, we provide details of our implementation (Section A.1), datasets (Section A.2), and experimental results (Section A.3). We will also open-source our Python code to help researchers replicate our experiments.

## A.1 Implementation Details

We now describe implementation details for each component in CLTS: seed word extraction in $L_S$, seed word transfer, and teacher-student co-training in $L_T$.

**Source Seed Word Extraction:** The inputs to the classifier in $L_S$ are tf-idf weighted unigram vectors[7]. For the classifier, we use scikit-learn's logistic regression[8] with the following parameters: penalty="l1", C=$\lambda_B$, solver="liblinear", multi_class="ovr". In other words, we address multi-class classification by training $K$ binary "one-vs.-rest" logistic regression classifiers to minimize the $L1$-regularized logistic loss (LASSO). (We use scikit-learn version 0.22.1, which does not support a "multinomial" loss with L1-penalized classifiers.) We tune $\lambda_B$ by computing the "regularization path" between 0.1 and $10^7$, evenly spaced on a log scale into 50 steps. To efficiently[9] compute the regularization path, we use the "warm-start" technique (Koh et al., 2007), where the solution of the previous optimization step is used to initialize the solution for the next one. This is supported in scikit-learn by setting the warm_start parameter of logistic regression to True.

**Seed Word Transfer:** We obtain seed-word translations using the MUSE[10] bilingual dictionaries (Lample et al., 2018), which contain up to 100,000 dictionary entries per language pair. Importantly, we use only the translations for $B \leq 500$ English seed words. To understand the impact of translation budget in performance, we experiment with the following values for $\frac{B}{K}$: [2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200]. We leave for future work the non-uniform distribution of seed words across classes, which might improve efficiency as "easier" classes may be modeled with fewer seed words. For Uyghur and Sinhalese, which have no entries in MUSE, we use Google Translate. For reproducibility, we cached the translations obtained from Google Translate and will share them with the code of the paper. If a source word has multiple translations in MUSE,[11] we use all translations as noisy target seed words with the same weight, while if a seed word has no translation in the target language, then we directly use it as a target seed word (this may be useful for named entities, emojis, etc.). Translations provided by a human annotator would possibly lead to better target seed words but, as we show here, even noisy automatic translations can be effectively used in CLTS.

**Teacher-Student Co-Training:** For the logistic regression (LogReg) student in $L_T$, we use scikit-learn's logistic regression with default parameters (including penalty="l2", C=1). The inputs to LogReg are tf-idf weighted n-gram ($n$=1,2) vectors. For our monolingual BERT (MonoBERT) student, we use the following pre-trained models from huggingface[12]:

- English: bert-base-cased
- Spanish: dccuchile/bert-base-spanish-wwm-cased
- French: camembert-base
- German: bert-base-german-cased
- Italian: dbmdz/bert-base-italian-xxl-cased
- Russian: DeepPavlov/rubert-base-cased
- Chinese: bert-base-chinese
- Japanese: bert-base-japanese

We use the default hyperparameters in the "Transformers" library (Wolf et al., 2019) and do not re-train (with the language modeling objective) MonoBERT in the target domain. To avoid label distribution shift because of iterative co-training, we balance teacher-labeled documents in $D'_T$ by keeping the same number of documents across classes before training the student. We perform

[7] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[9] Using a 16-core CPU machine, we compute $\lambda_B$ and train the source classifier in less than one minute (see Section A.3).

[10] https://github.com/facebookresearch/MUSE#ground-truth-bilingual-dictionaries

[11] Various translations for a word in MUSE may correspond to different senses of the word. For example, the seed word "shares" for the "Corporate" topic translates to both "comparte" (share) and "acciones" (stocks) in Spanish.

[12] https://huggingface.co

two rounds of teacher-student co-training, which has been shown to gain most of the improvement in Karamanolakis et al. (2019). Table 11 reports the model parameters for each dataset and language. We do not tune any model hyperparameters and use default values instead.

### A.2 Dataset Details

**Document Classification in MLDoc:** The Multilingual Document Classification Corpus (MLDoc[13]; Schwenk and Li (2018)) contains Reuters news documents in English, German, Spanish, French, Italian, Russian, Chinese, and Japanese. Each document is labeled with one of the four categories:

- CCAT (Corporate/Industrial)
- ECAT (Economics)
- GCAT (Government/Social)
- MCAT (Markets)

MLDoc was pre-processed and split by Schwenk and Li (2018) into 1,000 training, 1,000 validation, and 4,000 test documents for each language (Table 1). We use labeled training documents only in English for training the source classifier. We treat training documents in German, Spanish, French, Italian, Russian, Chinese, and Japanese as unlabeled in CLTS by ignoring the labels.

**Review Sentiment Classification in CLS:** The Cross-Lingual Sentiment corpus (CLS[14]; Prettenhofer and Stein (2010)) contains Amazon product reviews in English, German, French, and Japanese. Each language includes product reviews from three domains: books, dvd, and music. Each labeled document includes a binary (positive, negative) sentiment label. Table 2 reports dataset statistics. Validation sets are not available for CLS. We use labeled training documents only in English for training the source classifier. We ignore training documents in German, French, and Japanese, and use unlabeled documents in CLTS.

**Sentiment Classification in TwitterSent, Sentipers, and LORELEI:** The Twitter Sentiment corpus (TwitterSent; Mozetič et al. (2016)) contains Twitter posts in Bulgarian (Bg), German (De), English (En), Spanish (Es), Croatian (Hr), Hungarian (Hu), Polish (Pl), Portuguese (Pt), Slovak (Sk), Slovenian (Sl), and Swedish (Sv). We use the pre-processed and tokenized data provided by (Rasooli et al., 2018). In addition to these tweets, Rasooli et al. (2018) also use pre-processed and tokenized Persian (Fa) product reviews from the SentiPers corpus (Hosseini et al., 2018) and manually labeled Uyghur (Ug) documents from the LDC LORELEI corpus. On the above datasets, each document is labeled with a sentiment label: positive, neutral, or negative. Table 3 reports dataset statistics. We use labeled training documents only in English for training the source classifier. We treat training documents in the rest of the languages as unlabeled.

| Language | Train | Dev | Test |
|---|---|---|---|
| English (En) | 1,000 | 1,000 | 4,000 |
| German (De) | 1,000 | 1,000 | 4,000 |
| Spanish (Es) | 1,000 | 1,000 | 4,000 |
| French(Fr) | 1,000 | 1,000 | 4,000 |
| Italian (It) | 1,000 | 1,000 | 4,000 |
| Russian (Ru) | 1,000 | 1,000 | 4,000 |
| Chinese (Zh) | 1,000 | 1,000 | 4,000 |
| Japanese (Ja) | 1,000 | 1,000 | 4,000 |

Table 1: MLDoc corpus statistics.

| Language | Domain | Train | Unlabeled | Test |
|---|---|---|---|---|
| English | books | 2,000 | 10,000 | 2,000 |
| | dvd | 2,000 | 10,000 | 2,000 |
| | music | 2,000 | 10,000 | 2,000 |
| German | books | 2,000 | 30,000 | 2,000 |
| | dvd | 2,000 | 30,000 | 2,000 |
| | music | 2,000 | 30,000 | 2,000 |
| French | books | 2,000 | 30,000 | 2,000 |
| | dvd | 2,000 | 16,000 | 2,000 |
| | music | 2,000 | 30,000 | 2,000 |
| Japanese | books | 2,000 | 30,000 | 2,000 |
| | dvd | 2,000 | 9,000 | 2,000 |
| | music | 2,000 | 30,000 | 2,000 |

Table 2: CLS corpus statistics.

**Medical Emergency Situation Classification in LORELEI:** The Low Resource Languages for Emergent Incidents (LORELEI) corpus (Strassel and Tracey, 2016) contains (among others) documents in Uyghur (Ug)[15] and Sinhalese (Si)[16]. Each document is labeled with an emergency need. Similar to Yuan et al. (2020), we consider binary classi-

[13] https://github.com/facebookresearch/MLDoc

[14] https://webis.de/data/webis-cls-10.html

[15] LDC2016E57_LORELEI_Uyghur

[16] LDC2018E57_LORELEI_Sinhalese

| Language | Train | Dev | Test |
|---|---|---|---|
| Arabic | - | 671 | 6100 |
| Bulgarian | 23985 | 2999 | 2958 |
| German | 63748 | 7970 | 7961 |
| English | 46645 | 5832 | 5828 |
| Spanish | 137205 | 17152 | 17133 |
| Persian | 15000 | 1000 | 3027 |
| Croatian | 56368 | 7047 | 7025 |
| Hungarian | 36224 | 4528 | 4520 |
| Polish | 116241 | 14531 | 14517 |
| Portuguese | 63082 | 7886 | 7872 |
| Russian | 44780 | 5598 | 5594 |
| Slovak | 40476 | 5060 | 5058 |
| Slovenian | 74268 | 9285 | 9277 |
| Swedish | 32601 | 4076 | 4074 |
| Uyghur | - | 136 | 346 |

Table 3: Twittersent, SentiPers, and LORELEI corpus statistics.

fication to medical versus non-medical emergency need. Unfortunately, our number of labeled documents for each language is different than that reported in Yuan et al. (2020). In English, we use 806 labeled documents for training the source classifier. In Uyghur, we use 5,000 unlabeled documents for training the student and 226 labeled documents for evaluation. In Sinhalese, we use 5,000 unlabeled documents for training the student and 36 labeled documents for evaluation. Given the limited number of labeled documents, we do not consider validation sets for our experiments.

### A.3 Experimental Result Details

We now discuss detailed results on each dataset. In addition to baselines reported in the main paper, we also report supervised classifiers (*-sup) that were trained on each language separately using the labeled training data, to get an estimate for the maximum achievable performance. We run CLTS 5 times using the following random seeds: [7, 20, 42, 127, 1993] and report the average performance results and the standard deviation across different runs. (The standard deviation for our LogReg student is negligible across all datasets so we do not report it.) We report the results for the configuration of $B$ that achieves the best validation performance (accuracy for MLDoc, macro-average F1 for TwitterSent) and also report the validation performance, when a validation set is available.

Table 6 reports results on MLDoc. Eisenschlos et al. (2019) report two different results for LASER (Artetxe and Schwenk, 2019): LASER-paper are the results reported in (Artetxe and Schwenk, 2019), while LASER-code are different results using the most recent LASER code. Here, we report both. (In Table 4a, we have reported the LASER configuration that achieves the best performance for each language.) As expected, the performance of supervised models that consider in-language training datasets is higher than cross-lingual models.

Table 7 reports results on CLS per domain. (In Table 4b, we reported the average performance across domains for each language.) Note that MultiFiT-sup has substantially higher accuracy than MonoBERT-sup and LogReg-sup. This indicates that MulfiFit is probably a better model for this task. It would be interesting to evaluate in the future whether using MultiFiT as student outperforms Student-MonoBERT.

Table 8 reports results on TwitterSent, SentiPers, and LORELEI. We have reported the best performing approaches in Rasooli et al. (2018) that use En as a source language. We noticed that CLTS achieves best validation performance using more seed words in the Twitter corpora compared to the MLDoc and CLS corpora. We hypothesize that because Twitter posts are shorter than news documents or reviews, the context of seed words is less rich in indicative words and so the student requires larger teacher-labeled datasets to be effective. Note, however, that even with a tighter budget of $B$=60, CLTS-Student has an average accuracy of 40.5% and outperforms previous approaches relying on dictionaries or comparable corpora.

**Examples of Extracted Seed Words:** Table 4 reports the 10 most important seed words extracted for each of the four news document classes in CLS. Table 5 reports the 10 most important seed words extracted for each binary class and domain in CLS. Figure 8 reports the 20 most important seed words extracted for each of the 3 sentiment classes in TwitterSent, SentiPers and LORELEI. Figure 9 reports the 20 most important seed words extracted for the medical situation class in LORELEI and their translations to Uyghur and Sinhalese.

**Testing CLTS in Non-English Source Languages:** To evaluate whether our results generalize to non-English source languages, we run additional experiments using De, Es, and Fr as source languages in CLS. For those experiments, we also consider En as a target language. Table 9 reports the evaluation results. Across all configurations, there is no clear winner between MultiCCA and MultiBERT, but our Student-LogReg consistently outperforms both approaches, indicating that CLTS is also effective with non-English source languages.

**Ablation Study:** Table 10 reports results on MLDoc by changing parts of CLTS. The first row reports Student-Logreg without any changes. **Change (a):** using the clarity-scoring (similar to tf-idf weighting) method of (Angelidis and Lapata, 2018) leads to 3% lower accuracy than extracting seed words from the weights of a classifier trained through sparsity regularization. **Change (b):** obtaining translations through Google Translate leads to 0.8% lower accuracy than using bilingual MUSE dictionary. We observed that Google Translate sometimes translates words to wrong translations without extra context, while MUSE dictionaries provide more accurate translations. **Change (c):** updating Teacher similar to Karamanolakis et al. (2019), where the Teacher updates seed word qualities but does not consider documents without seed words during training, leads to 1.3% lower accuracy than our approach, which replaces the teacher by the student and thus considers even documents without seed words. **Change (d):** removing seed words from Student's input leads to 2.8% lower accuracy than letting Student consider both seed words and non-seed words. This shows that even without using seed words, Student still performs accurately (77.2% accuracy across languages), indicating that Student successfully exploits indicative features in the context of the seed words.

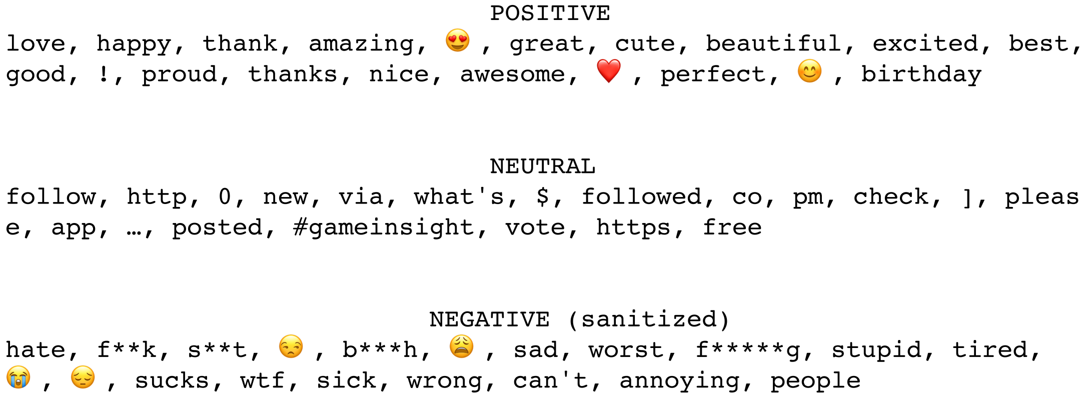
**Runtime:** Table 12 reports the end-to-end runtime for each experiment (i.e., the total time needed to run the script), which includes: loading data, training, and evaluating CLTS. The runtime does not include dataset pre-processing, which was performed only once. We ran all experiments on a server with the following specifications: 16 CPUs, RAM: 188G, main disk: SSD 1T, storage disk: SDD 3T, GPU: Titan RTX 24G.

| | |
|---|---|
| CCAT | company, inc, ltd, corp, group, profit, executive, newsroom, rating, shares |
| ECAT | bonds, economic, deficit, inflation, growth, tax, economy, percent, foreign, budget |
| GCAT | president, police, stories, party, sunday, people, opposition, beat, win, team |
| MCAT | traders, futures, dealers, market, bids, points, trading, day, copper, prices |

Table 4: MLDoc: Top 10 English seed words extracted per class (Section 4.1).

| | |
|---|---|
| DVD-POS | best, great, excellent, love, highly, enjoy, wonderful, life, good, favorite |
| BOOK-POS | excellent, great, lives, wonderful, life, fascinating, fun, easy, love, best |
| MUSIC-POS | amazing, highly, great, favorites, best, favorite, awesome, classic, excellent, love |
| DVD-NEG | waste, boring, worst, bad, disappointing, disappointed, awful, poor, horrible, terrible |
| BOOKS-NEG | money, disappointed, disappointing, boring, disappointment, worst, waste, bad, finish, terrible |
| MUSIC-NEG | boring, worst, disappointment, poor, sorry, garbage, money, disappointing, bad, horrible |

Table 5: CLS: Top 10 English seed words extracted per class and domain (Section 4.1).



POSITIVE

love, happy, thank, amazing, 😍 , great, cute, beautiful, excited, best,
good, !, proud, thanks, nice, awesome, ❤️ , perfect, 😊 , birthday

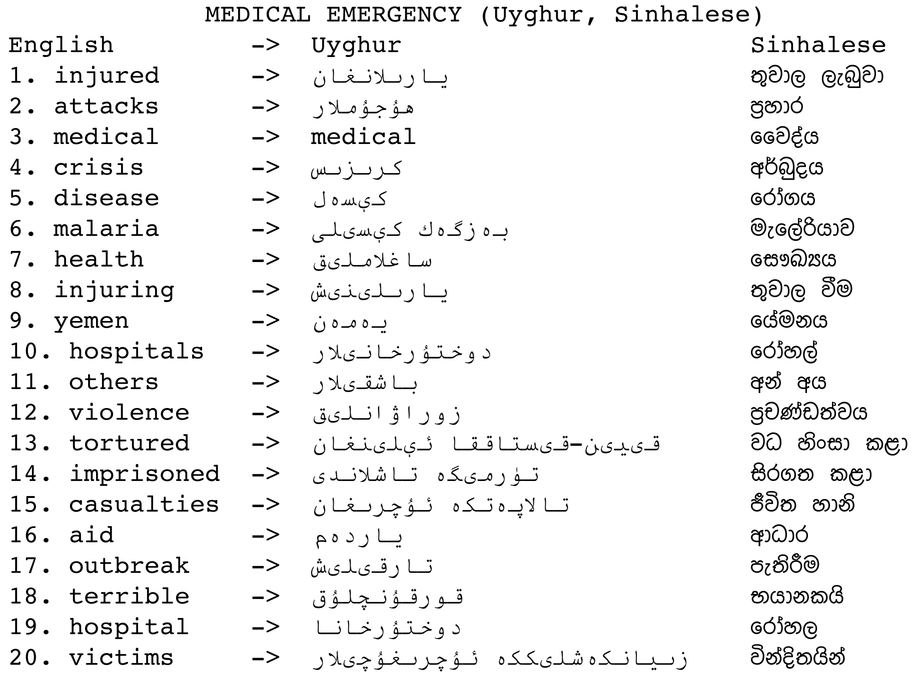
NEUTRAL

follow, http, 0, new, via, what's, $, followed, co, pm, check, ], pleas
e, app, …, posted, #gameinsight, vote, https, free

NEGATIVE (sanitized)

hate, f**k, s**t, 😒 , b***h, 😩 , sad, worst, f*****g, stupid, tired,
😭 , 😔 , sucks, wtf, sick, wrong, can't, annoying, people

Figure 8: TwitterSent: Top 20 seed words extracted per class (Section 4.1). Interestingly, some of the seed words are actually not words but emojis used by Twitter users to indicate the corresponding sentiment class.



MEDICAL EMERGENCY (Uyghur, Sinhalese)

| English | -> | Uyghur | Sinhalese |
|---|---|---|---|
| 1. injured | -> | يارىلانغان | තුවාල ලැබුවා |
| 2. attacks | -> | ھۇجۇملار | ප්‍රහාර |
| 3. medical | -> | medical | වෛද්‍ය |
| 4. crisis | -> | كرىزىس | අර්බුදය |
| 5. disease | -> | كېسەل | රෝගය |
| 6. malaria | -> | بەزگەك كېسىلى | මැලේරියාව |
| 7. health | -> | ساغلاملىق | සෞඛ්‍යය |
| 8. injuring | -> | يارىلىنىش | තුවාල වීම |
| 9. yemen | -> | يەمەن | යේමනය |
| 10. hospitals | -> | دوختۇرخانىلار | රෝහල් |
| 11. others | -> | باشقىلار | අන් අය |
| 12. violence | -> | زوراۋانلىق | ප්‍රචණ්ඩත්වය |
| 13. tortured | -> | قىيىن-قىستاقتا ئېلىنغان | වධ හිංසා කළා |
| 14. imprisoned | -> | تۈرمىگە تاشلاندى | සිරගත කළා |
| 15. casualties | -> | تالاپەتكە ئۇچرىغان | ජීවිත හානි |
| 16. aid | -> | ياردەم | ආධාර |
| 17. outbreak | -> | تارقىلىش | පැතිරීම |
| 18. terrible | -> | قورقۇنچلۇق | භයානකයි |
| 19. hospital | -> | دوختۇرخانا | රෝහල |
| 20. victims | -> | زىيانكەشلىككە ئۇچرىغۇچىلار | වින්දිතයින් |

Figure 9: LORELEI: Top 20 seed words for the "medical emergency" class and their translations obtained through Google Translate. The incorrect translation for the important "medical" seed word from English to Uyghur is "medical."

| **Method** | **Cross-Lingual** | **Language** | | | | | | | **AVG** |
|---|---|---|---|---|---|---|---|---|---|
| | **Resource** | **De** | **Es** | **Fr** | **It** | **Ru** | **Zh** | **Ja** | **Acc** |
| *Methods below use labeled target documents (supervised)* | | | | | | | | | |
| LogReg-sup | - | 93.7 | 93.8 | 91.6 | 85.2 | 83.7 | 87.6 | 88.4 | 89.1 |
| MultiBERT-sup | - | 93.3 | 95.7 | 93.4 | 88.0 | 87.5 | 89.3 | 88.4 | 90.8 |
| MultiFiT-sup | - | **95.9** | **96.1** | **94.8** | **90.3** | **87.7** | **92.6** | **90.0** | **92.5** |
| *Methods below use parallel corpora* | | | | | | | | | |
| LASER, paper | parallel corpora | 86.3 | **79.3** | 78.3 | 70.2 | 67.3 | 71.0 | 61.0 | 71.2 |
| LASER, code | parallel corpora | 87.7 | 75.5 | 84.0 | 71.2 | 66.6 | 76.7 | 64.6 | 75.2 |
| MultiFiT | LASER | **91.6** | 79.1 | **89.4** | **76.0** | **67.8** | **82.5** | **69.6** | **79.4** |
| *Methods below use pre-trained multi-lingual language models* | | | | | | | | | |
| MultiBERT | - | 79.8 | 72.1 | 73.5 | 63.7 | 73.7 | 76.0 | 72.8 | 73.1 |
| ST-MultiBERT | MultiBERT | **90.0** | **85.3** | **88.4** | **75.2** | **79.3** | **87.0** | **76.8** | **83.1** |
| *Methods below use bilingual dictionaries (Student requires Teacher)* | | | | | | | | | |
| MultiCCA | $B = 20K$ | 81.2 | 72.5 | 72.4 | 69.4 | 60.8 | 74.7 | 67.6 | 71.2 |
| Teacher | MUSE ($B = 160$) | 72.7 | 73.5 | 77.6 | 62.5 | 46.9 | 53.3 | 31.9 | 59.8 |
| Student-LogReg | Teacher | 87.4 | 86.0 | 89.1 | 70.5 | 71.9 | 82.4 | 68.8 | 79.4 |
| Student-MonoBERT | Teacher | **90.4** | **86.3** | **91.2** | **74.7** | **75.6** | **84.0** | **72.6** | **82.1** |
| *Below we report the standard deviation of test accuracies across 5 runs* | | | | | | | | | |
| Student-MonoBERT | | 0.5 | 0.5 | 0.4 | 0.7 | 0.8 | 0.4 | 0.6 | |
| *Below we report validation accuracies* | | | | | | | | | |
| Teacher | MUSE ($B = 160$) | 72.9 | 74.1 | 79.5 | 59.5 | 54.8 | 65.7 | 49.0 | |
| Student-LogReg | Teacher | 86.5 | 88.4 | 88.5 | 70.9 | 73.2 | 82.3 | 67.7 | |
| Student-MonoBERT | Teacher | 89.8 | 88.2 | 91.6 | 75.2 | 76.9 | 84.2 | 71.1 | |

Table 6: Accuracy results on MLDoc.

| **Method** | **Cross-Lingual** | **De** | | | **Fr** | | | **Ja** | | | **AVG** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Resource** | Books | DVD | Music | Books | DVD | Music | Books | DVD | Music | **Acc** |
| *Methods below use labeled target documents (supervised)* | | | | | | | | | | | |
| LogReg-sup | - | 84.5 | 82.8 | 84.1 | 84.7 | 86.0 | 88.0 | 80.9 | 83.0 | 83.0 | 84.1 |
| MultiBERT-sup | - | 86.1 | 84.1 | 82.0 | 86.2 | 86.9 | 86.7 | 80.9 | 82.8 | 80.0 | 84.0 |
| MonoBERT-sup | - | 82.4 | 80.0 | 81.7 | 88.4 | 86.2 | 86.3 | **86.3** | 85.7 | 86.2 | 84.8 |
| MultiFiT-sup | - | **93.2** | **90.5** | **93.0** | **91.3** | **89.6** | **93.4** | **86.3** | **85.8** | **86.6** | **90.0** |
| *Methods below use parallel corpora or MT systems* | | | | | | | | | | | |
| MT-BOW | GoogleTransl. | 79.7 | 77.9 | 77.2 | 80.8 | 78.8 | 75.8 | 70.2 | 71.3 | 72.0 | 76.0 |
| BiDRL | Google Transl. | 84.4 | **84.1** | **84.7** | 84.4 | **83.6** | 82.5 | 73.2 | 76.8 | 78.8 | 81.3 |
| CLDFA | parallel corpora | 84.0 | 83.1 | 79.0 | 83.4 | 82.6 | 83.3 | 77.4 | **80.5** | 76.5 | 81.1 |
| LASER, code | parallel corpora | 84.2 | 78.0 | 79.2 | 83.9 | 83.4 | 80.8 | 75.0 | 75.6 | 76.3 | 79.5 |
| MultiFiT | LASER | **89.6** | 81.8 | 84.4 | **87.8** | 83.5 | **85.6** | **80.5** | 77.7 | **81.5** | **83.6** |
| *Methods below use bilingual dictionaries or no cross-lingual systems* | | | | | | | | | | | |
| VECMAP | - | 76.0 | 76.3 | 73.5 | 77.8 | 78.6 | 78.1 | 55.9 | 57.6 | 54.4 | 69.8 |
| MultiBERT | - | 72.2 | 70.1 | 73.8 | 75.5 | 74.7 | 76.1 | 65.4 | 64.9 | 70.3 | 71.4 |
| CL-SCL | $B = 450$ pivots | 79.5 | 76.9 | 77.8 | 78.5 | 78.8 | 77.9 | 73.1 | 71.1 | 75.1 | 76.5 |
| Teacher | MUSE ($B = 20$) | 42.1 | 36.0 | 36.3 | 47.9 | 51.6 | 46.2 | 17.9 | 23.9 | 26.2 | 36.5 |
| Student-LogReg | Teacher | 76.0 | 77.8 | 82.2 | 78.8 | 80.0 | 80.1 | **77.2** | **79.8** | **78.9** | 79.0 |
| Student-MonoBERT | Teacher | 77.9 | **79.9** | **82.5** | **84.3** | 83.9 | **82.0** | 76.4 | 77.7 | 78.8 | **80.4** |
| *Below we report the standard deviation of test accuracies across 5 runs* | | | | | | | | | | | |
| Student-MonoBERT | Teacher | 0.6 | 0.9 | 0.8 | 0.9 | 0.4 | 0.5 | 0.5 | 0.4 | 0.2 | |

Table 7: Accuracy results on CLS. Validation accuracy is not reported as there is no validation set.

| Method | CL Resource | **Ar** | **Bg** | **De** | **Es** | **Fa** | **Hr** | **Hu** | **Pl** | **Pt** | **Ru** | **Sk** | **Sl** | **Sv** | **Ug** | **AVG** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Methods below labeled target documents (supervised)* | | | | | | | | | | | | | | | | |
| LogReg-sup | - | - | 54.4 | 54.4 | 43.8 | 65.9 | 56.0 | 51.4 | 56.4 | 49.9 | 56.6 | 66.0 | 57.0 | 60.6 | - | - (56.0) |
| LSTM-sup | - | - | 54.5 | 59.9 | 45.4 | 67.8 | 61.6 | 60.4 | 64.5 | 51.1 | 69.2 | 70.1 | 58.6 | 62.5 | - | - (60.5) |
| *Methods below use parallel corpora* | | | | | | | | | | | | | | | | |
| CLWE-BQ | parallel corpora | 37.3 | 33.0 | 43.5 | 42.6 | 40.1 | 30.8 | 41.1 | 41.7 | 38.6 | 44.8 | 22.6 | 32.2 | 39.1 | 30.0 | 37.0 |
| Project-LDC | parallel corpora | 37.2 | - | - | 42.7 | 33.1 | - | 47.0 | - | - | 48.0 | - | | - | 38.6 | - (41.1) |
| Project-EP | parallel corpora | - | **38.7** | **47.3** | 41.8 | - | - | 38.1 | 38.8 | **39.3** | - | **30.0** | **44.6** | **44.6** | - | (40.4) |
| *Methods below use comparable corpora or dictionaries* | | | | | | | | | | | | | | | | |
| SentiWordNet | SentiWordNet (>20K) | 25.6 | 30.6 | 32.0 | 25.3 | 25.3 | 19.8 | 29.2 | 26.0 | 22.9 | 29.5 | 19.2 | 28.1 | 22.7 | 36.7 | 26.6 |
| CLWE-Wikt | Wiktionary (>20K) | 31.0 | 45.3 | 51.0 | 37.7 | 31.7 | - | 40.8 | 32.9 | 35.4 | **43.8** | 36.6 | 32.1 | 40.4 | 28.0 | - (37.4) |
| CLWE-CP | comparable corpora | 21.1 | 28.6 | 37.7 | 27.7 | 20.7 | 13.9 | 22.4 | 30.2 | 22.2 | 25.3 | 24.6 | 25.3 | 31.1 | 25.7 | 25.5 |
| Teacher | $B = 500$ | 22.7 | 42.8 | 45.5 | 42.7 | 30.9 | 36.4 | 39.4 | 40.7 | 34.4 | 29.8 | 40.4 | 29.5 | 38.7 | 20.3 | 35.3 |
| Student-LogReg | Teacher | **39.0** | **46.3** | **52.5** | **44.9** | **45.7** | **39.4** | **45.2** | **45.4** | 38.7 | 43.2 | **43.3** | **42.1** | **50.4** | **41.2** | **44.1** |
| *Below we report validation accuracies* | | | | | | | | | | | | | | | | |
| Teacher | $B = 500$ | 31.3 | 43.8 | 45.7 | 43.2 | 32.3 | 34.3 | 39.4 | 41.1 | 35.0 | 27.2 | 40.5 | 29.7 | 40.5 | 22.8 | |
| Student-LogReg | Teacher | 47.2 | 48.7 | 52.1 | 45.4 | 46.5 | 39.0 | 46.9 | 45.3 | 40.1 | 41.7 | 43.2 | 42.5 | 50.0 | 38.3 | |

Table 8: Macro-averaged F1 results on TwitterSent, SentiPers, and LDC LORELEI.

| | **Target Acc** (MultiCCA / MultiBERT / Student-LogReg) | | | |
|---|---|---|---|---|
| Source Language | En | De | Es | Fr |
| En | - | 81.2/80.2/**87.4** | 72.5/76.9/**86.0** | 72.4/72.6/**89.1** |
| De | 56.0/59.7/**82.8** | - | 73.2/54.0/**81.3** | 71.6/60.0/**84.9** |
| Es | 74.0/74.2/**80.8** | 55.8/57.6/**83.3** | - | 65.6/71.8/**89.0** |
| Fr | 64.8/76.1/**84.1** | 53.7/51.8/**84.5** | 65.4/72.1/**85.5** | - |

Table 9: MultiCCA (left) vs. MultiBERT (center) vs. Student-LogReg (right) for various train (rows) and test (columns) configurations on MLDoc. Student-LogReg substantially outperforms MultiCCA and MultiBERT across all train and test configurations: CLTS effectively transfers weak supervision also from non-English source languages.

| **Change** | **AVG Acc** |
|---|---|
| - (Original Student-LogReg) | **79.4** |
| (a) Extract seed words as in Angelidis and Lapata (2018) | 77.0 (↓ 3.0%) |
| (b) Replace MUSE translations by Google Translate | 78.8 (↓ 0.8%) |
| (c) Update Teacher as in Karamanolakis et al. (2019) | 78.4 (↓ 1.3%) |
| (d) Remove seed words from Student's input | 77.2 (↓ 2.8%) |

Table 10: Ablation experiments on MLDoc.

| Dataset | Lang | LogReg | MonoBERT |
|---|---|---|---|
| MLdoc | De | 14104 | 109M |
|  | Es | 15080 | 110M |
|  | Fr | 17632 | 111M |
|  | It | 11676 | 111M |
|  | Ru | 26804 | 178M |
|  | Zh | 15248 | 102M |
|  | Ja | 24676 | 111M |
| CLS-books | De | 37560 | 109M |
|  | Fr | 33462 | 111M |
|  | Ja | 67195 | 111M |
| CLS-dvd | De | 49832 | 109M |
|  | Fr | 12448 | 111M |
|  | Ja | 61897 | 111M |
| CLS-music | De | 49899 | 109M |
|  | Fr | 27194 | 111M |
|  | Ja | 60554 | 111M |
| TwitterSent | Ar | 5502 | - |
|  | Bg | 44565 | - |
|  | De | 105993 | - |
|  | Es | 245778 | - |
|  | Fa | 44811 | - |
|  | Hr | 108030 | - |
|  | Hu | 50532 | - |
|  | Pl | 184266 | - |
|  | Pt | 83685 | - |
|  | Ru | 58416 | - |
|  | Sk | 76776 | - |
|  | Sl | 140226 | - |
|  | Sv | 70902 | - |
|  | Ug | 978 | - |
| LORELEI | Ug | 1353 | - |
|  | Si | 4654 | - |

Table 11: Number of model parameters for our LogReg and MonoBERT student in each dataset and language.

| Dataset | Lang | LogReg | MonoBERT |
|---|---|---|---|
| MLdoc | De | 61s | 176s |
|  | Es | 33s | 165s |
|  | Fr | 43s | 139s |
|  | It | 29s | 157s |
|  | Ru | 54s | 195s |
|  | Zh | 70s | 173s |
|  | Ja | 51s | 170s |
|  | AVG | 49s | 168s |
| CLS-books | De | 247s | 699s |
|  | Fr | 301s | 837s |
|  | Ja | 256s | 785s |
| CLS-dvd | De | 158s | 641s |
|  | Fr | 71s | 277s |
|  | Ja | 125s | 317s |
| CLS-music | De | 272s | 925s |
|  | Fr | 290s | 884s |
|  | Ja | 238s | 800s |
|  | AVG | 218s | 685s |
| TwitterSent | Ar | 32s | - |
|  | Bg | 82s | - |
|  | De | 367s | - |
|  | Es | 2176s | - |
|  | Fa | 60s | - |
|  | Hr | 282s | - |
|  | Hu | 120s | - |
|  | Pl | 1445s | - |
|  | Pt | 361s | - |
|  | Ru | 164s | - |
|  | Sk | 181s | - |
|  | Sl | 654s | - |
|  | Sv | 145s | - |
|  | Ug | 20s | - |
|  | AVG | 434s | - |
| LORELEI | Ug | 160s | - |
|  | Si | 174s | - |
|  | AVG | 167s | - |

Table 12: Runtimes for our LogReg and MonoBERT student in each dataset and language.