LAMM: Language Aware Active Learning for Multilingual Models

Ze Ye*
yeze@amazon.com
Amazon.com, Inc.
Seattle, WA, USA

Dantong Liu*
lidanton@amazon.com
Amazon.com, Inc.
Sunnyvale, CA, USA

Kaushik Pavani
sripava@amazon.com
Amazon.com, Inc.
Seattle, WA, USA

Sunny Dasgupta
sunnyd@amazon.com
Amazon.com, Inc.
Seattle, WA, USA

ABSTRACT

In industrial settings, it is often necessary to achieve language-level accuracy targets. For example, Amazon business teams need to build multilingual product classifiers that operate accurately in all European languages. It is unacceptable for the accuracy of product classification to meet the target in one language (e.g, English), while falling below the target in other languages (e.g, Portuguese). To fix such issues, we propose Language Aware Active Learning for Multilingual Models (LAMM), an active learning strategy that enables a classifier to learn from a small amount of labeled data in a targeted manner to improve the accuracy of Low-resource languages (LRLs) with limited amounts of data for model training. Our empirical results on two open-source datasets and two proprietary product classification datasets demonstrate that LAMM is able to improve the LRL performance by 4%–11% when compared to strong baselines.

ACM Reference Format:
Ze Ye, Dantong Liu, Kaushik Pavani, and Sunny Dasgupta. 2023. LAMM: Language Aware Active Learning for Multilingual Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, Birmingham, UK, 2 pages. https://doi.org/10.1145/3583780.3615507

1 INTRODUCTION

Cross-lingual learning is known to improve the accuracy of NLP models on Low-Resource Languages (LRLs), that have limited amounts of data for model training, by leveraging shared representations that are common to multiple languages. However, even state-of-the-art (SOTA) models (e.g., mBERT, XLM-R) have a sizable gap in the accuracy of high-resource languages (HRLs) and LRLs, which correlates to the relative scarcity of pre-training data from the LRLs [1]. Such problem is also common in Amazon, as evidenced by our experiments on multi-lingual product classification. This can be explained by the skewed distribution across languages in the Amazon catalog where there is significantly fewer Portuguese product data than French product data.

In industrial settings, we are often required to achieve language-level accuracy targets (e.g., 0.95 accuracy for each language). In such cases, it is a common practice to synthetically augment the existing training data by translating using neural machine translators, transliterating, or label propagation on similar data [2, 3]. These techniques can help increase the size and diversity of training data without involving expensive human annotation. While translation typically helps increase LRL accuracy, we may not achieve business targets due to the inherent imperfections of machine translation.

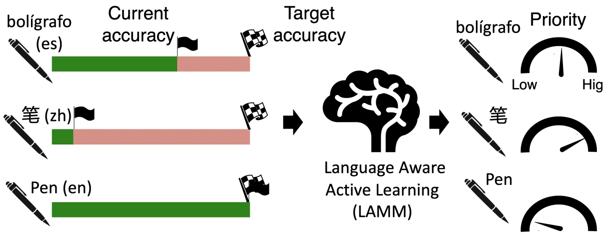

Figure 1: Consider a multilingual pen classifier (to classify whether a object is a pen or not) which comprises 3 languages (English, Spanish, and Chinese), LAMM computes relative priority weights for each language based on its current estimated accuracy (solid flag) and target accuracy (checked flag). In this example, the expected priority is Chinese > Spanish > English.

We present a method for increase low-resource language (LRL) accuracy by utilizing active learning to obtain labeled data in a targeted manner. Previous research on active learning for multilingual data mainly concentrated on training a single model for a single language [4–6]. Our goal, however, is to develop one model that performs well across multiple languages. The most similar prior research to our approach, Language Aware Active Learning for Multilingual Models (LAMM), is the work of [7], which demonstrated that active learning can boost model performance for multiple languages when using a single model. Unfortunately, this method does not provide a way to target a LRL and improve its accuracy. In scenarios where reaching language-level targets is necessary, current state-of-the-art active learning algorithms fall short, as they persist in obtaining labels for languages that have already met their accuracy objectives, thus wasting manual annotations. To address this issue, we introduce an active-learning-based approach for obtaining labeled data in a targeted fashion to enhance LRL accuracy without compromising high-resource language (HRL) performance. LAMM, our proposed method, raises the chances of meeting accuracy goals for all languages of interest.

2 LAMM

We pose LAMM as a multi-objective optimization problem. The goal is to select unlabeled data instances that are both (1) uncertain (i.e., the model is not confident in its predictions), and (2) from languages for which the classifier achieves lower accuracy than our targets. Specifically, we decide if each unlabeled data instance x_j (j denotes the index of the data point in unlabeled dataset) is selected

*Co-first authors with equal contribution

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615507

for labeling with two measures: (1) uncertainty in prediction scores H_j, (2) a language-specific weighting parameter $L_{j'}$ (j' denotes the language that data point j belongs to, $j' \in J'$, where J' is the set of languages contained in unlabeled data and we assume the language of each data point is known). In one active learning iteration, we select batch $\hat{B}$ of unlabeled data with B elements as:

$$\hat{B} = \underset{S' \subset S, |S'|=B}{\operatorname{argmax}} \sum_{s \in S'} s_j \tag{1}$$

Here S is a list of scores determining if an unlabeled data needs to be selected. The score $s_j \in S$ is a function of the uncertainty score H_j and the language weight $L_{j'}$. $s_j = f(H_j, L_{j'})$. $f(.)$ can be altered for different use cases, as long as $f(.)$ has positive correlation with H_j, and $L_{j'}$. $|S'|$ denotes the number of elements in the subset S'. Overall, the mathematical expression in Eq. (1) denotes a subset of elements $\hat{B}$ with size B that have the highest sum of the function $f(.)$.

In our implementation, H_j is the entropy of classification probability scores calculated as $H(s) = -\sum_{i=1}^{k} p_i \log p_i$, where k is the number of possible labels in a multi-class classifier, p_i is the probability that the classifier assigns to label i, and log is the natural logarithm. The language weight $L_{j'}$ is decided based on the performance gap between the currently achieved accuracy of each language and its accuracy target. The higher gap results in higher language weight. More specifically, we divide the overall annotation budget (B) in each active learning iteration among different languages based on their language weight $L_{j'}$. Within each language's budget, we select the samples with high uncertainty H_j from that language. It is worth mentioning although in this work we implement LAMM on top of the entropy based query strategy, LAMM can be applied to any active learning strategy to create its language-aware variant.

3 EXPERIMENT

We employ the standard pool-based active learning set up, and evaluate the performance of LAMM against 2 baselines on 4 multi-lingual classification datasets. Amazon Review[8] and MLDoc[9] are two public datasets. Dataset A and B are two Amazon internal multilingual product classification datasets. The baseline methods are: (1) Least Confidence (LC) which acquires samples with the highest uncertainty score (entropy), (2) Equal Allocation (EA) where annotation budget is equally allocated to different languages and then samples with high entropy in respective language are acquired to fill per-language annotation budget. In experiments, we set the classification accuracy target for each language as 0.95 and 0.90 for the Amazon internal datasets and public datasets, respectively. As machine translation is commonly used to increase multi-lingual model performance, the experiments are conducted with machine translation, which means after the query strategy acquires the samples, we use AWS Translate to translate the acquired samples to all the other languages-of-interest, and then train model with all the data. Table 1 shows the detailed language level performance of each dataset. In all the subsequent result analysis, for the Dataset A, Dataset B and Amazon Review datasets, we define HRL as the language with the largest proportion in the pool data, and LRLs as all the other languages in each dataset. From Table 1 we note LAMM achieves the best performance across LRLs compared to other baselines while only takes a small performance hit on HRL. We notice that LAMM reduces the proportion of HRL labels by 62.1% (from avg. 70.35% to 26.67%) while only reducing HRL's accuracy AUC by 1.2% (from avg. 0.752 to 0.743) compared with LC.

Table 1: Language level performance for all datasets at last active learning iteration (with 1000 labeled data) in the scenario with translation. L is the language in each dataset, $\mathcal{D}$ (%) is data proportion of each language in the pool set, $\mathcal{L}$ (%) is the labeled data proportion of each language, and $\mathcal{A}$ (%) is area under accuracy curve (AUC) of each language. For each dataset, we highlight highest AUC number in bold.Accuracy AUC is the area under each accuracy curve (AUC) summarizing accuracy across all label counts/active learning iterations.

Dataset	L	$\mathcal{D}$	LC		LAMM		EA	
			$\mathcal{L}$	$\mathcal{A}$	$\mathcal{L}$	$\mathcal{A}$	$\mathcal{L}$	$\mathcal{A}$
Dataset A	fr	78	74.7	**0.772**	36.3	0.756	37.2	0.754
	pt	12	10	0.709	40.7	**0.722**	31.3	0.714
	ja	10	15.2	0.766	22.9	**0.78**	31.4	0.78
Dataset B	en	77	59.7	**0.749**	12.4	0.746	26.4	0.748
	fr	9.4	19.7	0.682	41.3	**0.683**	18.9	0.674
	de	6.2	8.8	0.675	16.5	**0.692**	18.1	0.679
	ja	3.5	7	0.707	10.4	**0.714**	18.3	0.714
	pt	3.3	4.7	0.732	19.1	0.733	18.3	**0.739**
Amazon Review	en	73	72.1	0.713	25.4	**0.714**	30.9	0.702
	fr	18	4.8	0.686	21.1	**0.698**	22.7	0.688
	de	4.5	18.4	0.667	24.4	**0.683**	24.1	0.680
	ja	4.5	4.7	0.658	29.1	**0.664**	22.3	0.655
MLDoc	high	73	74.7	**0.774**	32.6	0.757	31.0	0.762
	mid	18	17.7	0.760	26.2	**0.761**	25.2	0.760
	low	9	7.6	0.748	41.2	**0.766**	43.8	0.761

REFERENCES

[1] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online, November 2020.

[2] Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Proc. Interspeech 2021*, pages 1529–1533, 2021.

[3] Gözde Gül Sahin and Mark Steedman. Data augmentation via dependency tree morphing for low-resource languages. *CoRR*, 2019.

[4] Longhua Qian, Haotian Hui, Ya'nan Hu, Guodong Zhou, and Qiaoming Zhu. Bilingual active learning for relation classification via pseudo parallel corpora. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 582–592, Baltimore, Maryland, June 2014.

[5] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. *CoRR*, abs/1707.05928, 2017.

[6] Manuel Tonneau, Dhaval Adjodah, Joao Palotti, Nir Grinberg, and Samuel Fraiberger. Multilingual detection of personal employment status on Twitter. In *Proceedings of the 60th Annual Meeting of the ACL*, Dublin, Ireland, May 2022.

[7] Joel Ruben Antony Moniz, Barun Patra, and Matthew Gormley. On efficiently acquiring annotations for multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, May 2022.

[8] Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[9] Holger Schwenk and Xian Li. A Corpus for Multilingual Document Classification in Eight Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 7-12, 2018 2018.