Self-Training Pre-Trained Language Models for Zero- and Few-Shot Multi-Dialectal Arabic Sequence Labeling

Muhammad Khalifa 1, Muhammad Abdul-Mageed2, Khaled Shaalan3
1Department of Computer Science, Cairo University
2Natural Language Processing Lab, University of British Columbia
3The British University in Dubai
m.khalifa@grad.fci-cu.edu.eg
muhammad.mageed@ubc.ca, khaled.shaalan@buid.ac.ae

Abstract
A sufficient amount of annotated data is usually required to fine-tune pre-trained language models for downstream tasks. Unfortunately, attaining labeled data can be costly, especially for multiple language varieties and dialects. We propose to self-train pre-trained language models in zero- and few-shot scenarios to improve performance on data-scarce varieties using only resources from data-rich ones. We demonstrate the utility of our approach in the context of Arabic sequence labeling by using a language model fine-tuned on Modern Standard Arabic (MSA) only to predict named entities (NE) and part-of-speech (POS) tags on several dialectal Arabic (DA) varieties. We show that self-training is indeed powerful, improving zero-shot MSA-to-DA transfer by as large as \( \sim 10\% \) F1 (NER) and 2% accuracy (POS tagging). We acquire even better performance in few-shot scenarios with limited amounts of labeled data. We conduct an ablation study and show that the performance boost observed directly results from the unlabeled DA examples used for self-training. Our work opens up opportunities for developing DA models exploiting only MSA resources and it can be extended to other languages and tasks.

1 Introduction
Neural language models (Xu and Rudnicky, 2000; Bengio et al., 2003) with vectorized word representations (Mikolov et al., 2013) are currently core to a very wide variety of NLP tasks. In specific, using representations from transformer-based (Vaswani et al., 2017) language models (Devlin et al., 2018; Liu et al., 2019), pre-trained on large amounts of unlabeled data and then fine-tuned on labeled task-specific data, has become a popular approach for improving downstream task performance. This pre-training then fine-tuning scheme has been successfully applied to several tasks, including question answering (Yang et al., 2019), social meaning detection (Abdul-Mageed et al., 2020d), text classification (Liu et al., 2019), named entity recognition (NER), and part-of-speech (POS) tagging (Tsai et al., 2019; Conneau et al., 2019). The same setup also works well for cross-lingual learning (Lample and Conneau, 2019; Conneau et al., 2019).

Given that it is very expensive to glean labeled resources for all language varieties and dialects, a question arises: “How can we leverage resource-rich dialects to develop models nuanced to downstream tasks for resource-scarce ones?”. In this work, we aim to answer this particular question by applying self-training to unlabeled target dialect data. We empirically show that self-training is indeed an effective strategy in zero-shot (where no gold dialectal data are included in training set, Section 4.2) and few-shot (where a given number of gold dialectal data points is included in training split, Section 4.4).

Our few-shot experiments reveal that self-training is always a useful strategy that consistently improves over mere fine-tuning, even when all dialect-specific gold data are used for fine-tuning. In order to understand why this is the case (i.e., why combining self-training with fine-tuning yields better results than mere fine-tuning), we perform an extensive error analysis based on our NER data. We discover that self-training helps the model most (\( % = 59.7\) with improving false positives. This includes DA tokens whose MSA orthographic counterparts (Shaalan, 2014) are either named entities or trigger words that frequently co-occur with named entities in MSA. Interestingly, such out-of-MSA tokens occur in highly dialectal contexts (e.g., interjections and idiomatic expressions employed in interpersonal social media communication) or ones.
where the social media context in which the language (DA) is employed affords more freedom of speech (Alshehri et al., 2020) and a platform for political satire. We present our error analysis in Section 5.

**Context:** Language use in social media tends to diverge from ‘standard’, offline norms (Danet and Herring, 2007; Herring et al., 2015). For example, users employ slang, emojis, abbreviations, letter repetitions, and other types of playful practices. This poses a challenge for processing social media data in general. However, there are other challenges specific to Arabic that motivate our work. More specifically, we choose Arabic to apply our approach since it affords a rich context of linguistic variation: In addition to the standard variety, MSA, Arabic also has several spoken dialects (Abdul-Mageed et al., 2018; Bouamor et al., 2019; Abdul-Mageed et al., 2020b,c), which differ significantly from the written MSA (Zaidan and Callison-Burch, 2014) thus offering an excellent context for studying our problem. Arabic dialects differ among themselves and from MSA at various linguistic levels: lexical, phonological, morphological, and syntactic. This makes our case much more challenging than that of standard vs. social media English, for example. For a good zero-shot performance in our case, a model is required to accommodate not only lexical distance between MSA and DA, but also differences in word formation and syntax (related to POS tags, for example) and lexical ambiguity (as the meaning of the same token can vary cross-dialectically). This makes the zero-shot setting even harder, where the performance drops 20% F1 points (See section 4.2).

From a geopolitical perspective, Arabic also has a strategic significance. This is a function of Arabic being the native tongue of 400 million speakers in 22 countries, spanning across two continents (Africa and Asia)\(^2\). In addition, the three dialects of our choice, namely Egyptian (EGY), Gulf (GLF), and Levantine (LEV), are popular dialects that are widely used online. This makes our resulting models highly useful in practical situations at scale. Pragmatically, ability to develop NLP systems on dialectal tasks with no-to-small labeled dialect data immediately eases a serious bottleneck. Arabic dialects differ among themselves and from MSA at all linguistic levels, posing challenges to traditional NLP approaches. We also note that our method is language-independent, and we hypothesize it can be directly applied to other varieties of Arabic or in other linguistic contexts for other languages and varieties.

**Tasks:** We apply our methods on two sequence labeling tasks, where we have access to both MSA and DA gold data. In particular, as mentioned above, we perform experiments on POS tagging and NER. Each of these tasks has become an integral part of various other NLP applications, including question answering, aspect-based sentiment analysis, machine translation, and summarization, and hence our developed models should have wide practical use. Again, we note that our approach itself is task-independent. The same approach can thus be applied to other tasks involving DA. We leave testing our approach on other languages, varieties, and tasks for future research.

**Contributions:** Our work offers the following contributions:

1. We study the problem of MSA-to-DA transfer in the context of sequence labeling and show that when training on MSA data only, a wide performance gap exists between testing on MSA and DA. That is, models fine-tuned on MSA generalize poorly to DA in zero-shot settings.

2. We propose self-training to improve zero- and few-shot MSA-to-DA transfer. Our approach requires little-to-no labeled DA data. We evaluate extensively on 3 different dialects, and show that our method indeed narrows the performance gap between MSA and DA by a margin as wide as \(\leq 10\%\) F1 points.

3. We develop state-of-the-art models for the two sequence labeling tasks (NER and POS).

We now introduce our method.

2 **Method**

While the majority of labeled Arabic datasets are in MSA, most daily communication in the Arab world is carried out in DA. In this work, we show that models trained on MSA for NER and POS tagging generalize poorly to dialect inputs when used in zero-shot-settings (i.e., no dialect data used during training). Across the two tasks, we test how self-training would fare as an approach to leverage unlabeled DA data to improve performance
on DA. As for self-training, it involves training a model using its own predictions on a set of unlabeled data identical from its original training split. Our proposed self-training procedure is given two sets of examples: a labeled set \( L \) and an unlabeled set \( U \). To perform zero-shot MSA-to-DA transfer, MSA examples are used as the labeled set, while unlabeled DA examples are the unlabeled set. As shown in Figure 1, each iteration of the self-training algorithm consists mainly of three steps. First, a pre-trained language model is fine-tuned on the labeled MSA examples \( L \). Second, for every unlabeled DA example \( u_i \), we use the model to tag each of its tokens to obtain a set of predictions and confidence scores for each token \( p_{ui} = (l_1^{(i)}, c_1^{(i)}), (l_2^{(i)}, c_2^{(i)}), \ldots, (l_{|u_i|}^{(i)}, c_{|u_i|}^{(i)}) \), where \( (l_j^{(i)}, c_j^{(i)}) \) are the label and confidence score (softmax probability) for the \( j \)-th token in \( u_i \). Third, we employ a selection mechanism to identify examples from \( U \) that are going to be added to \( L \) for the next iteration.

For a selection mechanism, we experiment with both a thresholding approach and a fixed-size (Dong and de Melo, 2019) approach. In the thresholding method, a threshold \( \tau \) is applied on the minimum confidence per example. That is, we only add an example \( u_i \) to \( L \) if \( \min_{(l_j^{(i)}, c_j^{(i)}) \in p_{ui}} c_j^{(i)} \geq \tau \). See Algorithm 1. The fixed-size approach involves, at each iteration, the selection of the top \( S \) examples with respect to the minimum confidence score \( \min_{(l_j^{(i)}, c_j^{(i)}) \in p_{ui}} c_j^{(i)} \), where \( S \) is a hyper-parameter. We experiment with both approaches and report results in Section 4.

For our language model, we use XLM-RoBERTa (Conneau et al., 2019), XLM-R for short. XLM-R is a cross-lingual model, and we choose it since it is reported to perform better than the multilingual mBERT (Devlin et al., 2018). XLM-R also uses Common Crawl for training, which is more likely to have dialectal data than the Arabic Wikipedia (used in mBERT), making it more suited to our work. We now introduce our experiments.

3 Experiments

We begin our experiments with evaluating the standard fine-tuning performance of XLM-R models on both NER and POS tagging against strong baselines. We then use our best models from this first round to investigate the MSA-to-DA zero-shot transfer, showing a significant performance drop even when using pre-trained XLM-R. Consequently, we employ self-training for both NER and POS tagging in zero- and few-shot settings, showing substantial performance improvements in both cases. We now introduce our datasets.

3.1 Datasets

NER: For our work on NER, we use 4 datasets: ANERcorp (Benajiba et al., 2007), ~150K tokens; ACE 2003 (Mitchell et al., 2003) BNews (BN-2003), ~15K tokens; ACE 2003 Newswire (NW-2003), ~27K tokens; and Twitter (Darwish, 2013), ~81K tokens. Named entity types in all datasets are location (LOC), organization (ORG), and person (PER).

POS Tagging: There are a number of Arabic POS tagging datasets, mostly on MSA (Maamouri et al., 2004) but also on dialects such as EGY (Maamouri et al., 2014). To show that the proposed approach is able to work across multiple dialects, we ideally needed data from more than one dialect. Hence, we use the multi-dialectal

---

**Algorithm 1: MSA-to-DA Self-Training**

1. **Given** set \( L \) of labeled MSA examples, set \( U \) of unlabeled DA examples, \( \tau \) parameter for probability threshold selection.
2. **repeat**
   3. Fine-tune model \( M \) for \( K \) epochs on labeled MSA examples \( L \);
   4. for \( u_i \in U \) do
   5. Observe prediction \( p_{ui} \) on unlabeled DA example \( u_i \) using model \( M \);
   6. if \( \min_{(l_j^{(i)}, c_j^{(i)}) \in p_{ui}} c_j^{(i)} \geq \tau \) then
   7. remove \( u_i \) from \( U \) and add it to \( L \);
   8. end
3. **until** stopping criterion satisfied

---

![Figure 1: MSA-to-DA Self-training transfer.](image-url)
dataset from (Darwish et al., 2018), comprising 350 tweets from each of the 4 varieties MSA, EGY, GLF and LEV. This dataset has 21 POS tags, some of which are suited to social media (since it is derived from Twitter). We show the POS tag set from (Darwish et al., 2018) in Table 10 in Appendix A. We now introduce our baselines.

3.2 Baselines

For the NER task, we use the following baselines:

- **NERA (Abdallah et al., 2012)**: A hybrid system of rule-based features and a decision tree classifier.

- **WC-BiLSTM (Gridach, 2016)**: A character- and a word-level Bi-LSTM with a conditional random fields (CRF) layer.

- **WC-CNN (Khalifa and Shaalan, 2019)**: A character- and a word-level CNN with a CRF layer.

- **mBERT (Devlin et al., 2018)**: A fine-tuned multilingual BERT-Base-Cased (110M parameters), pre-trained with a masked language modeling objective on the Wikipedia corpus of 104 languages (including Arabic). For fine-tuning, we find that (based on experiments on our development set) a learning rate of $6 \times 10^{-5}$ works best with a dropout of 0.1.

In addition, we compare to the published results in (Shaalan and Oudah, 2014), AraBERT (Antoun et al., 2020), and CAMel (Obeid et al., 2020) for the ANERCorp dataset. We also compare to the published results in (Khalifa and Shaalan, 2019) for the 4 datasets.

For the POS tagging task, we compare to our own implementation of WC-BiLSTM (since there is no published research that uses this method on the task, as far as we know) and run mBERT on our data. We also compare to the CRF results published by (Darwish et al., 2018). In addition, for the Gulf dialect, we compare to the BiLSTM with compositional character representation and word representations (CC2W+W) published results in (Alharbi et al., 2018).

3.3 Experimental Setup

Our main models are XLM-RoBERTa base architecture XLM-RB ($L = 12, H = 768, A = 12, 270M$ params) and XLM-RoBERTa large architecture XLM-RL ($L = 24, H = 1024, A = 16, 550M$ params), where $L$ is number of layers, $H$ is the hidden size, $A$ is the number of self-attention heads. For XLM-R experiments, we use Adam optimizer with $1e^{-5}$ learning rate, batch size of 16. We typically fine-tune for 20 epochs, keeping the best model on the development set for testing. We report results on the test split for each dataset, across the two tasks. For all BiLSTM experiments, we use the same hyper-parameters as (Khalifa and Shaalan, 2019).

For the standard fine-tuning experiments, we use the same train/development/test split as in (Khalifa and Shaalan, 2019) for NER, and the same split provided by (Darwish et al., 2018) for POS tagging. For all the self-training experiments, we use the dialect subset of the Arabic online news commentary (AOC) dataset (Zaidan and Callison-Burch, 2011), comprising the EGY, GLF, and LEV varieties limiting to equal sizes of 9K examples per dialect (total =27K). We use the split from (Elaraby and Abdul-Mageed, 2018) of AOC, removing the dialect labels and just using the comments themselves for our self-training. Each iteration involved fine-tuning the model for $K = 5$ epochs. As a stopping criterion, we use early stopping with patience of 10 epochs. Other hyper-parameters are set as listed before. For selecting confident samples, we experiment with a fixed number of top samples $S = [50, 100, 200]$ and selection based on a probability threshold $\tau = [0.80, 0.90, 0.95]$ (softmax values)\(^4\). For all evaluations, we use the seqeval toolkit\(^5\).

4 Results

4.1 Fine-tuning XLM-R

Here, we show the results of standard fine-tuning of XLM-R for the two tasks in question. We start by showing the result of fine-tuning XLM-R on the named entity task, on each of the 4 Arabic NER (ANER) datasets listed in Section 3.1. Table 1 shows the test set macro $F_1$ score on each of the 4 ANER datasets. Clearly, the fine-tuned XLM-R models outperform other baselines on all datasets,\(^6\)

\(^4\)We note that our approach could be scaled with an even bigger unlabeled dataset, given the performance gains we report with self-training in this work.

\(^5\)It is worth noting that our $S$ values are similar to those used in (Dong and de Melo, 2019). We also experimented with other values for $\tau$ and $S$, but found them sub-optimal and hence we report performance only for the listed values of these two hyper-parameters here.

\(^6\)https://github.com/chakki-works/seqeval.
except on the NW-2003 where WC-CNN (Khalifa and Shaalan, 2019) performs slightly better than XLM-R_{L}.

For POS Tagging, Table 2 shows test set word accuracy of the XLM-R models compared to baselines. Again, XLM-R models (both base and large) outperform all other models. A question arises why XLM-R models outperform both mBERT and AraBERT. As noted before, for XLM-R vs. mBERT, XLM-R was trained on much larger data: CommonCrawl for XLM-R vs. Wikipedia for mBERT. Hence, the larger dataset of XLM-R is giving it an advantage over mBERT. For comparison with AraBERT, although the pre-training data for XLM-R and AraBERT may be comparable, even the smaller XLM-R model (XLM-R_{B}) has more than twice the number of parameters of the BERT_{BASE} architecture on which AraBERT and mBERT are built (270M v. 110M). Hence, XLM-R model capacity gives it another advantage. We now report our experiments with zero-shot transfer from MSA to DA.

4.2 MSA-DA Zero-Shot Transfer

We start by the discussion of NER experiments. Since there is no publicly available purely dialectal NER dataset on which we can study MSA-to-DA transfer, we needed to find DA data to evaluate on. We observed that the dataset from (Darwish, 2013) contains both MSA and DA examples (tweets). Hence, we train a binary classifier\(^6\) to distinguish DA data from MSA. We then extract examples that are labeled with probability \(p > 0.90\) as either DA or MSA. We obtain 2,027 MSA examples (henceforth, Twitter-MSA) and 1,695 DA examples (henceforth, Twitter-DA), respectively. We split these into development and test sets with 30% and 70% ratios. As for POS Tagging, we already have the three previously used DA datasets, namely EGY, GLF and LEV. We use those for the zero-shot setting by omitting their training sets and using only the development and test sets.

We first study how well models trained for NER and POS tagging on MSA data only will generalize to DA inputs during test time. We evaluate this zero-shot performance on both the XLM-R_{B} and XLM-R_{L} models. For NER, we train on ANERCorp (which is pure MSA) and evaluate on both Twitter-MSA and Twitter-DA. While for POS tagging, we train on the MSA subset (Darwish et al., 2018) and evaluate on the corresponding test set for each dialect. As shown in Table 3, for NER, a significant generalization gap of around 20% \(F_{1}\) points exists between evaluation on MSA and DA using both models. While for POS tagging, the gap is as large as 18.13% accuracy for the LEV dialect with XLM-R_{B}. The smallest generalization gap is on the GLF variety, which is perhaps due to the high overlap between GLF and MSA (Alharbi et al., 2018). In the next section, we evaluate the ability of self-training to close this MSA-DA performance gap.

4.3 Zero-shot Self-Training

Here, for NER, similar to Section 4.2, we train on ANERCorp (pure MSA) and evaluate on Twitter-MSA and Twitter-DA. Table 4 shows self-training NER results employing the selection mechanisms listed in Section 2, and with different values for \(S\) and \(\tau\). The best improvement is achieved with the thresholding selection mechanism with a \(\tau = 0.90\), where we have an \(F_{1}\) gain of 10.03 points. More generally, self-training improves zero-shot performance in all cases albeit with different \(F_{1}\) gains. It is noteworthy, however, that the much higher-capacity large model deteriorates on MSA if self-trained (dropping from 68.32% to 67.21%). This shows the ability of the large model to learn representations very specific to DA when self-trained. It is also interesting to see that the best self-trained base model achieved 50.10% \(F_{1}\), outperforming the large model before the latter is self-trained (47.35% in the zero-shot setting). As such, we conclude that a base self-trained model, with less computational capacity, can (and in our case does) improve over a large (not-self-trained) model that needs significant computation. The fact that, when self-trained, the large model improves 15.35% points over the base model in the zero-shot setting (55.42 vs. 40.07) is remarkable.

As for POS tagging, we similarly observe consistent improvements in zero-shot transfer with self-training (Table 5). The best model achieves accuracy gains of 2.41% (EGY), 1.41% (GLF), and 1.74% (LEV). Again, this demonstrates the utility of self-training pre-trained language models on the POS tagging task even in absence of labeled dialectal POS data (zero-shot).

---

\(^6\)The model we use is XLM-R_{B} fine-tuned on the AOC using (Elaraby and Abdul-Mageed, 2018) split. We achieve development and test accuracies of 90.3% and 89.4 %, respectively, outperforming the best results in (Elaraby and Abdul-Mageed, 2018).
| Model                                      | ANERCorp | BN-2003 | NW-2003 | Twitter |
|--------------------------------------------|----------|---------|---------|---------|
| NERA (Abdallah et al., 2012)              | 88.77    | –       | –       | –       |
| CAMeL (Obeid et al., 2020)                | 85.00    | –       | –       | –       |
| Hybrid (Shaalan and Oudah, 2014)          | 90.66    | –       | –       | –       |
| WC-BiLSTM (Gridach, 2016)                 | 88.56    | 94.92   | 90.32   | 64.93   |
| WC-CNN (Khalifa and Shaalan, 2019)        | 88.77    | 94.12   | 91.20   | 65.34   |
| mBERT (ours)                              | 85.86    | 89.52   | 87.19   | 58.92   |
| AraBERT (Antoun et al., 2020)             | 87.75    | 95.35   | 85.25   | 60.39   |
| XLM-R (ours)                              | 91.43    | 97.33   | 91.10   | 68.91   |

Table 1: Test set macro F\(_1\) scores for NER.

| Model                                      | MSA      | EGY      | GLF      | LEV      |
|--------------------------------------------|----------|----------|----------|----------|
| BiLSTM (CC2W + W) (Alharbi et al., 2018)  | 93.6     | 92.9     | 87.8     | 87.9     |
| CRF (Darwish et al., 2018)                 | 94.63    | 93.41    | 88.79    | 86.13    |
| WC-BiLSTM (ours)                           | 90.57    | 92.88    | 87.85    | 72.30    |
| mBERT (ours)                               | 96.30    | 94.70    | 92.18    | 89.98    |
| XLM-R\(_B\) (ours)                        | 98.21    | 97.00    | 94.41    | 93.19    |
| XLM-R\(_L\) (ours)                        | 91.43    | 97.33    | 91.10    | 68.91    |

Table 2: Test set accuracy for POS Tagging.

4.4 Few-Shot Self-Training

We also investigate whether self-training would be helpful in scenarios where we have access to some gold-labeled DA data (as is the case with POS tagging). Here, we evaluate the few-shot performance of self-training as increasing amounts of predicted DA data are added to the gold training set. This iteration of experiments focuses exclusively on POS tagging, using a fixed-size \( S = 100 \) of predicted cases for self-training and the XLM-R base model. Figure 2 shows how POS tagging test accuracy improves as the percentage of gold DA examples added to the MSA training data increases from 0% to 100% on the three dialects (EGY, GLF, and LEV). Comparing these results to those acquired via the standard fine-tuning settings without self-training, we find that self-training does consistently improve over fine-tuning. This improvement margin is largest with only 20% of the gold examples.

4.5 Ablation Study

Here, we conduct an ablation study with the NER task as our playground in order to verify our hypothesis that the performance boost primarily comes from using unlabeled DA data for self-training. By using a MSA dataset with the same size as our unlabeled DA one\(^7\), we can compare the performance of the self-trained model in both settings: MSA and DA unlabeled data. We run 3 different self-training experiments using 3 different values for \( \tau \) using each type of unlabeled data. Results are shown in table 6. While we find slight performance boost due to self-training even with MSA unlabeled data, the average F1 score with unlabeled DA is better by 2.67 points, showing that using unlabeled DA data for self-training has helped the model adapt to DA data during testing.

5 Error analysis

To understand why self-training the pre-trained language model, when combined with fine-tuning, improves over mere fine-tuning, we perform an error analysis. For the error analysis, we focus on the NER task where we observe a huge self-training gain. We use the development set of Twitter-DA (See section 4.3) for the error analysis. We compare predictions of the standard fine-tuned XLM-R\(_B\) model (FT) and the best performing self-training (\( \tau = 0.9 \)) model (ST) on the data, and provide the confusion matrices of both models with gold labels in Table 11 (in Appendix B). The error analysis leads to an interesting discovery: The greatest benefit from the ST model comes mostly from reducing false positives (see Table 7). In other words, self-training helps regularize the model predictions such that tokens misclassified by the original FT model as a named entities are now correctly tagged as unnamed entity “O”.

To understand why the ST model improves false positive rate, we manually inspect the cases it cor-
Table 3: Zero-shot transfer results on DA For NER (macro F1) and POS Tagging (accuracy). Models are trained on MSA only and evaluated on DA. Datasets used are: Twitter-MSA and Twitter-DA (Darwish, 2013) for NER, and Multi-dialectal (Darwish et al., 2018) for POS tagging.

| Model     | NER MSA | DA | POS MSA | EGY | GLF | LEV |
|-----------|---------|----|---------|-----|-----|-----|
| XLM-R_B   | 60.42   | 40.07 | 96.30   | 78.38 | 83.72 | 78.17 |
| XLM-R_L   | 68.32   | 47.35 | 98.21   | 82.28 | 85.95 | 81.24 |

Table 4: Zero-short self-training (ST) NER results. Models trained on ANERCorp (pure MSA) and evaluated on Twitter-MSA and Twitter-DA we extract from (Darwish et al., 2018). Self-training boosts the performance on DA data by 10% macro F1 points with XLM-R_B and $\tau = 0.90$.

| Model     | NER MSA | DA |
|-----------|---------|----|
| XLM-R_B   | 61.88   | 40.07 |
| XLM-R_B, ST, $S=50$ | 60.98   | 43.88 |
| XLM-R_B, ST, $S=100$ | 61.13   | 42.01 |
| XLM-R_B, ST, $S=200$ | 61.46   | 43.49 |
| XLM-R_B, ST, $\tau = 0.80$ | 63.36   | 46.97 |
| XLM-R_B, ST, $\tau = 0.90$ | 61.02   | 50.10 |
| XLM-R_B, ST, $\tau = 0.95$ | 62.25   | 47.91 |
| XLM-R_L   | 68.32   | 47.35 |
| XLM-R_L, ST, $\tau = 0.90$ | 67.21   | 55.42 |

Table 5: Zero-shot POS tagging transfer accuracy when training on MSA only. ST: self-training.

| Model     | NER MSA | EGY | GLF | LEV |
|-----------|---------|-----|-----|-----|
| XLM-R_B   | 96.30   | 78.38 | 83.72 | 78.17 |
| XLM-R_B, ST, $S=50$ | – | 80.79 | 85.13 | 79.91 |
| XLM-R_B, ST, $S=100$ | – | 80.43 | 84.74 | 79.16 |
| XLM-R_B, ST, $S=200$ | – | 78.75 | 84.21 | 79.40 |
| XLM-R_B, ST, $\tau = 0.80$ | – | 79.52 | 83.97 | 79.21 |
| XLM-R_B, ST, $\tau = 0.85$ | – | 78.97 | 83.53 | 79.06 |
| XLM-R_B, ST, $\tau = 0.90$ | – | 78.88 | 83.72 | 78.50 |
| XLM-R_L   | 98.21   | 82.28 | 85.95 | 81.24 |
| XLM-R_L, ST, $S=50$ | – | 82.65 | 87.76 | 83.70 |

Figure 2: Development accuracy as labeled DA data is added to the training MSA data. **Left:** Results on EGY. **Middle:** Results on GLF. **Right:** Results on LEV. We use fixed-size selection with $S = 100$ for self-training models.

We also investigate errors shared by both the FT and ST models (errors which the ST model also could not fix). Some of these errors result from the fact that often times both MSA and DA use the same word for both person and location names.
Table 6: Ablation experiment with MSA unlabeled data for zero-shot NER. Development set macro F1 is shown when using both unlabeled MSA and DA data with the same size. Average performance with DA unlabeled data is higher showing the effect of unlabeled DA on model final performance.

| Setting     | Unlabeled MSA | Unlabeled DA |
|-------------|---------------|--------------|
| ST, $\tau = 0.80$ | 43.88         | 44.46        |
| ST, $\tau = 0.90$ | 44.69         | 47.83        |
| ST, $\tau = 0.95$ | 43.43         | 46.87        |
| Avg         | 43.67         | 46.34        |

Table 7: Comparison of error categories in percentage between the fine-tuned model (FT) and the model combining fine-tuned+self-trained (ST) model, based on the dialectal part of the dev set of the NER task.

| Measure          | FT | ST | % improvement |
|------------------|----|----|---------------|
| True Positives   | 155| 165| 6.5%          |
| False Positive   | 159| 64 | 59.7%         |
| False Negative   | 162| 168| -3.7%         |
| True Negative    | 5,940| 6,035| 1.5%        |

Row (1) in Table 13 (in Appendix B) is an example where the word “Mubarak”, name of the ex-Egypt President, is used as LOC. Other errors include out-of-MSA tokens mistaken as named entities. An example is in row (3) in Table 13, where "" ("proof" or "basis" in EGY) is confused for "" ("emirate", which is a location). False trigger words, mentioned before, also play a role here. An example is in row (7) where "" is confused for PER due to the trigger word "" that is usually followed by a person name. Spelling mistakes cause the third source of errors, as in row (4). We also note that even with self-training, detecting ORG entities is more challenging than PER or LOC. The problem becomes harder when such organizations are not seen in training such as in rows (8) and (10) of Table 13.

6 Related Work

Sequence Labeling. Recent work on sequence labeling usually involves using a word- or character-level neural network with a CRF layer (Lample et al., 2016; Ma and Hovy, 2016). These architectures have also been applied to Arabic sequence tagging (Gridach, 2016; Alharbi et al., 2018; Khalifa and Shaalan, 2019; Al-Smadi et al., 2020; El Bazi and Laachfoubi, 2019), producing better or comparable results to classical rule-based approaches (Shaalan and Oudah, 2014). We refer the interested reader to (Shoufan and Alameri, 2015) and (Al-Ayyoub et al., 2018) for surveys on Arabic NLP.

Pre-trained Language Models. Language models, based on Transformers (Vaswani et al., 2017), and pre-trained with the masked language modeling (MLM) objective have seen wide use in various NLP tasks. Examples include BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), MASS (Song et al., 2019), and ELECTRA (Clark et al., 2020). While they have been applied to several tasks, including text classification, question answering, named entity recognition (Conneau et al., 2019), and POS tagging (Tsai et al., 2019), a sufficiently large amount of labeled data is required for good performance. Concurrent with our work, Abdul-Mageed et al. (2020a) released MARBERT, a language model trained on a large amount of dialectal Arabic data. However, the extent to which dialect-specific models such as MARBERT can alleviate lack of labeled data remains untested.

Cross-lingual Learning. Cross-lingual learning is of particular importance due to the scarcity of labeled resources in many of the world’s languages. The goal is to leverage existing labeled resources in high-resource languages (such as English) to optimize learning for low-resource ones. In our case, we leverage MSA resources for building DA models. With proximity to our work, Kim et al. (2017) trained a POS tagger for different languages...
Table 8: Sample false positives mitigated by self-training. These were correctly predicted as the unnamed entity “O” by the self-trained model.

| no. | Word               | Eng.                   | MSA          | Context/Explanation                                      | FT Pred. |
|-----|--------------------|------------------------|--------------|----------------------------------------------------------|----------|
| (1) | ندي               | we want                | ORG          | (we want to know who)                                     | PER      |
| (2) | ماكانوا             | wasn’t                 | PER          | (most of those who wasn’t believing)                      | LOC      |
| (3) | لوهولول              | LOL                    | PER          | (interjection)                                            | LOC      |
| (4) | عذان               | for                    | PER          | (she wants batteries to play)                              | PER      |
| (5) | دووفطي              | now                    | PER          | (convince him to move now)                                | PER      |
| (6) | ماذا                | what                   | PER          | (what do you think?)                                      | PER      |
| (7) | قادر               | capable                | PER          | (magically; idiomatic expression)                         | PER      |

Table 9: NER task. Sample false negatives produced by self-training.

| no. | Word               | Gold | FT | ST |
|-----|--------------------|------|----|----|
| (1) | الاحيان             | ORG  | ORG| O  |
| (2) | للبرادي             | PER  | PER| O  |
| (3) | عبد الافغاد          | PER  | PER| O  |
| (4) | فان ديرل             | PER  | PER| O  |
| (5) | الگرزو                 | PER  | PER| O  |
| (6) | زین پسون              | PER  | PER| O  |

using English-resources only using two BiLSTM networks to learn common and language-specific features. Xie et al. (2018) made use of bilingual word embeddings with self-attention to learn cross-lingual NER for low-resource languages.

Multilingual extensions of LMs have emerged through joint pre-training on multiple languages. Examples include mBERT (Devlin et al., 2018), XLM (Lample and Conneau, 2019) and XLM-RoBERTa (Conneau et al., 2019). Such multilingual models have become useful for few-shot and zero-shot cross-lingual settings, where there is little or no access to labeled data in the target language. For instance Conneau et al. (2019) evaluated a cross-lingual version of RoBERTa (Liu et al., 2019), namely XLM-R, on cross-lingual learning across different tasks such as question answering, text classification, and named entity recognition.

Self-Training. Self-Training is a semi-supervised technique to improve learning using unlabeled data. Self-training has been successfully applied to NER (Kozareva et al., 2005), POS tagging (Wang et al., 2007), parsing (Sagae, 2010) and text classification (Van Asch and Daelemans, 2016). Self-training has also been applied in cross-lingual settings when gold labels are rare in the target language. Hajmohammadi et al. (2015) proposed a combination of active learning and self-training for cross-lingual sentiment classification. Pan et al. (2017) made use of self-training for named entity tagging and linking across 282 different languages. Lastly, Dong and de Melo (2019) employed self-training to improve zero-shot cross-lingual classification with mBERT (Devlin et al., 2018).

7 Conclusion

Even though pre-trained language models have improved many NLP tasks, they still need labeled data for fine-tuning. We show how self-training can boost the performance of pre-trained language models in zero- and few-shot settings on various Arabic varieties. We apply our approach to two sequence labeling tasks (NER and POS), establishing new state-of-the-art results on both. Through in-depth error analysis and an ablation study, we uncover why our models work and where they can fail. Our method is language- and task-agnostic, and we believe it can be applied to other tasks and language settings. We intend to test this claim in future research. Our research also has bearings to ongoing work on language models and self-training, and interactions between these two areas can be the basis of future work. All our models and code are publicly available.

Acknowledgements

MAM gratefully acknowledges support from the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, Canadian Foundation for Innovation, Compute Canada (www.computecanada.ca) and UBC ARC-Sockeye (https://doi.org/10.14288/SOCKEYE).
References

Sherief Abdallah, Khaled Shaalan, and Mohammad Shoai. 2012. Integrating rule-based system with classification for Arabic named entity recognition. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 311–322. Springer.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020a. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. arXiv preprint arXiv:2011.01785.

Muhammad Abdul-Mageed, Chiyu Zhang, Houida Bouamor, and Nizar Habash. 2020b. Nadi 2020: The first nuanced arabic dialect identification shared task. In Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020), pages 97–110, Barcelona, Spain.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020c. Micro-dialect identification in diaphasic and code-switched environments. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5855–5876.

Muhammad Abdul-Mageed, Chiyu Zhang, Azadeh Hashemi, and El Moatez Billah Nagoudi. 2020d. AraNet: A Deep Learning Toolkit for Arabic Social Media. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 16–23.

Mahmoud Al-Ayyoub, Aya Nuseir, Khlooud Alsmeearat, Yaser Jararweh, and Brij Brij. 2018. Deep learning for arabic nlp: A survey. Journal of computational science, 26:522–531.

Mohammad Al-Smadi, Saad Al-Zboon, Yaser Jararweh, and Patrick Juola. 2020. Transfer learning for Arabic named entity recognition with deep neural networks. IEEE Access, 8:37736–37745.

Randah Alharbi, Walid Magdy, Kareem Darwish, Ahmed AbdelAli, and Hamdy Mubarak. 2018. Part-of-speech tagging for Arabic gulf dialect using bi-lstm. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Ali Alshehri, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2020. Understanding and detecting dangerous speech in social media. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 40–47, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for Arabic language understanding. arXiv preprint arXiv:2003.00104.

Yassine Benajiba, Paolo Rosso, and José Miguel Bened. 2007. Anersys: An Arabic named entity recognition system based on maximum entropy. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 143–153.

Joshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. Journal of machine learning research, 3(Feb):1137–1155.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 199–207.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

Brenda Danet and Susan C Herring. 2007. The multilingual Internet: Language, culture, and communication online. Oxford University Press.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1558–1567.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesokey, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect arabic pos tagging: a crf approach. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Xin Luna 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.
Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The penn Arabic treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an egyptian Arabic treebank: Impact of dialectal morphology on annotation and tool development. In *LREC*, pages 2348–2354.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, J Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstain, Lisa Ferro, and Beth Sundheim. 2003. Tides extraction (ace) 2003 multilingual training data. *LDC2004T09, Philadelphia, Penn.: Linguistic Data Consortium*.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhli Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Kenji Sagae. 2010. Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 37–44.

Khaled Shaalan. 2014. A survey of Arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510.

Khaled Shaalan and Mai Oudah. 2014. A hybrid approach to Arabic named entity recognition. *Journal of Information Science*, 40(1):67–87.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectal arabic: A survey. In *Proceedings of the second workshop on Arabic natural language processing*, pages 36–48.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Ariavazhagan, Xin Li, and Amelia Archer. 2019. Small and practical bert models for sequence labeling. *arXiv preprint arXiv:1909.00100*.
Vincent Van Asch and Walter Daelemans. 2016. Predicting the effectiveness of self-training: Application to sentiment classification. *arXiv preprint arXiv:1601.03288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wen Wang, Zhongqiang Huang, and Mary Harper. 2007. Semi-supervised learning for part-of-speech tagging of mandarin transcribed speech. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–137. IEEE.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. *arXiv preprint arXiv:1808.09861*.

Wei Xu and Alex Rudnicky. 2000. Can artificial neural networks learn language models? In *Sixth international conference on spoken language processing*.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357.

Omar Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.
Appendices

A POS Tag Set

Table 10 lists all the part-of-speech (POS) tags used in our experiments.

| Tag       | Description     | Tag       | Description     |
|-----------|-----------------|-----------|-----------------|
| ADV       | adverb          | ADJ       | adjective       |
| CONJ      | conjunction     | DET       | determiner      |
| NOUN      | noun            | NSUFF     | noun suffix     |
| NUM       | number          | PART      | particle        |
| PUNC      | punctuation     | PRON      | pronoun         |
| PREP      | preposition     | V         | verb            |
| ABBREV    | abbreviation    | VSUFF     | verb suffix     |
| FOREIGN   | non-Arabic      | FUT_PART  | future particle |
| PROG_PART | progressive particle | EMOT | Emoticon/Emoji |
| MENTION   | twitter mention | HASH      | Hashtag         |
| URL       | URL             | –         | –               |

Table 10: The POS tag set in (Darwish et al., 2018).

B Error Analysis
Table 11: NER confusion matrices for fine-tuning (left) and self-training (right) on the development set of the DA NER data.

| no. | Token | Context/Explanation | Gold | FT Pred. |
|-----|-------|---------------------|------|----------|
| (1) | نبي تعرف من... (we want to know who) | PER | PER |
| (2) | ما كانوا لم يكونوا (most of those who wasn’t believing) | LOC | PER |
| (3) | لولوول (interjection) | PER | PER |
| (4) | عشان تي بطاريات عشان للمش... (she wants batteries to play) | LOC | PER |
| (5) | الآن (convince him to move now) | LOC | PER |
| (6) | ماذا (what do you think?) | PER | PER |
| (7) | قادر (magically; idiomatic expression) | PER | PER |
| (8) | المشين طوطبي (shameful Tantawy; Playful for General Tant.) | PER | PER |
| (9) | أبديكم (I entreat you to convince him) | PER | PER |
| (10) | أسلال (I ask you what) | ORG | ORG |
| (11) | من (who do you vote for, Badawi) | PER | PER |
| (12) | قرص مرن (holds a floppy disk against him) | PER | PER |
| (13) | يطيب (I entertain you to convince him) | PER | PER |
| (14) | ماء (brought them water) | PER | PER |
| (15) | إعداد تغريد (if depressed click retweet) | PER | PER |

Table 12: NER task. Bigger sample false positives mitigated by self-training. These were correctly predicted as the unnamed entity “O” by the self-trained model.

| no. | Token(s) | Context/Explanation | Gold | FT | ST |
|-----|----------|---------------------|------|----|----|
| (1) | بالباز (We are still in Mubarak) | LOC | PER | O |
| (2) | دعاء (a drunk entered the lecture) | O | PER | PER |
| (3) | دعاء (what is the evidence/sign and where?) | O | LOC | LOC |
| (4) | Qasr AlDobara Hospital (to Qasr AlDobara Hospital) | LOC | O | O |
| (5) | كنتاني (by Kentucky restaurant) | LOC | O | O |
| (6) | مشروع داون تاون بطنطا (a down town Tanta project) | LOC | O | O |
| (7) | Congratulations, hero! (Congratulations, hero!) | O | PER | PER |
| (8) | يطيب (I entertain you to convince him) | PER | PER | PER |
| (9) | مشاهدة قناة العربية (watched Al Arabya Channel) | ORG | O | O |
| (10) | أو فلمند دوس ريتويت (if depressed click retweet) | ORG | O | O |

Table 13: NER task. Sample errors that are not fixed by self-training (shared with the mere fine-tuned model)