Exploiting Cross-Lingual Subword Similarities in Low-Resource Document Classification

Mozhi Zhang
CS and UMIACS
University of Maryland
College Park, MD
mozhi@umiacs.umd.edu

Yoshinari Fujinuma
Computer Science
University of Colorado
Boulder, CO
Yoshinari.Fujinuma@colorado.edu

Jordan Boyd-Graber
CS, iSchool, LSC, and UMIACS
University of Maryland
College Park, MD
jbg@umiacs.umd.edu

Abstract

Text classification must sometimes be applied in situations with no training data in a target language. However, training data may be available in a related language. We investigate whether character-level knowledge transfer from a related language helps text classification. We present a cross-lingual document classification framework (CACO) that exploits cross-lingual subword similarity by jointly training a character-based embedder and a word-based classifier. The embedder derives vector representations for input words from their written forms, and the classifier makes predictions based on the word vectors. We use a joint character representation for both the source language and the target language, which allows the embedder to generalize knowledge about source language words to target language words with similar forms. We propose a multi-task objective that can further improve the model if additional cross-lingual or monolingual resources are available. Experiments confirm that character-level knowledge transfer is more data-efficient than word-level transfer between related languages.

1 Introduction: Classifiers across Languages

Modern machine learning methods in natural language processing can learn highly accurate, context-based classifiers (Devlin et al., 2019; Peters et al., 2018). Despite this revolution for high-resource languages such as English, Chinese, or German, some languages are left behind because of the dearth of text data generally and specifically labeled data. Often, the need for fixing this disparity is acute: an earthquake in Haiti, unrest in Ukraine, or food shortages in East Africa. Cross-lingual document classification (Klementiev et al., 2012, CLDC) attacks this problem by using annotated dataset from a source language to build classifiers for a target language.

CLDC works when it can find a shared representation for documents from both languages: train a classifier on source language documents and apply it on target language documents. Previous work uses a bilingual lexicon (Shi et al., 2010; Andrade et al., 2015), machine translation (Banea et al., 2008; Wan, 2009; Zhou et al., 2016, MT), topic models (Mimno et al., 2009; Yuan et al., 2018), or pre-trained cross-lingual word embeddings (Klementiev et al., 2012; Chen et al., 2018, CLWE) to extract cross-lingual features. But these methods may be impossible in low-resource languages, as they require some combination of large parallel or comparable text, high-coverage dictionaries, and monolingual corpora from a shared domain.

However, as anyone who has puzzled out a Portuguese menu from their high school Spanish knows, the task is not hopeless, as languages do not exist in isolation. Shared linguistic roots, geographic proximity, and history bind languages together; cognates abound, words sound the same, and there are often shared morphological patterns. This paper investigates character-level knowledge transfer for CLDC in truly low-resource settings, where unlabeled or parallel data in the target language is also limited or unavailable.

To study knowledge transfer at character level, we propose a CLDC framework, Classification Aided by Convergent Orthography (CACO) that capitalizes on subword similarities between related language pairs. Previous CLDC methods treat words as atomic symbols and do not transfer subword patterns across languages; CACO instead uses a bi-level model with two components: a character-based embedder and a word-based classifier.
The embedder exploits shared patterns in related languages to create word representations from character sequences. The classifier can then use the shared representation across languages to label the document. We hope the embedder can learn morpho-semantic regularities, while the classifier connects lexical semantics to labels.

To allow cross-lingual transfer, we use a single model for both languages, and we share character embeddings between source and target languages. We jointly train the embedder and the classifier on annotated source language documents. The embedder transfers knowledge about source language words to target language words with similar orthographic features.

While the model works well without any target language data, it can also benefit from a small amount of additional information when available. If we have a dictionary, pre-trained monolingual word embeddings, or parallel text, we can fine-tune the accuracy of the model to match the output of the reference model in another high-resource language, we can train the embedder with a small amount of additional information when available.

We verify the effectiveness of character-level knowledge transfer on two CLDC benchmarks. When we have enough data to learn high-quality CLWE, training classifiers with CLWE as input features is a strong CLDC baseline. CACO can match the accuracy of CLWE-based models without using any target language data, and fine-tuning the embedder with a small amount of additional resources improves CACO’s accuracy. Finally, CACO is still useful when we have enough resources to train high-quality CLWE—using CLWE as extra features, CACO outperforms the baseline CLWE-based models.

2 Classification Aided by Convergent Orthography

This section introduces our method, CACO, which trains a multilingual document classifier using labeled datasets in a source language $S$ and applies the classifier to a low-resource target language $T$.

2.1 Model Architecture

Our model has a two-level architecture (Figure 1) that includes a character-level embedder $e$ and a word-level classifier $f$. The embedder $e$ takes characters as inputs and produces a word embedding vector. The classifier then computes a label distribution from the embeddings of all input words. Formally, let $w$ be an input document with a sequence of tokens $w = \langle w_1, w_2, \ldots, w_n \rangle$. Our model maps the document $w$ to a distribution over possible labels $y$ in two steps. First, we generate a word embedding $v_i$ for each input word $w_i$ using the embedder $e$:

$$v_i = e(w_i).$$

We then feed the word embeddings to the classifier $f$ to compute the distribution over labels $y$:

$$p(y | w) = f(v_1, v_2, \ldots, v_n).$$

We can use any sequence model for the embedder $e$ and the classifier $f$. For our experiments, we use a bidirectional LSTM (Graves and Schmidhuber, 2005, BI-LSTM) embedder and a deep averaging network (Iyyer et al., 2015, DAN) classifier.

**BI-LSTM Embedder.** BI-LSTM is a powerful sequence model that captures complex non-local dependencies. Character-level BI-LSTM embedders are successfully used in many natural language processing tasks (Ling et al., 2015; Ballesteros et al., 2015; Lample et al., 2016). To embed a word $w$, we pass its character sequence $c$ to a left-to-right LSTM and the reversed character sequence $c^r$ to a right-to-left LSTM. We concatenate the final hidden states of the two LSTM and apply a linear transformation:

$$c(w) = W_e \cdot [\text{LSTM}(c) ; \text{LSTM}(c^r)] + b_e,$$

where the functions $\text{LSTM}$ and $\text{LSTM}$ compute the final hidden states of the two LSTMs.

**DAN Classifier.** A DAN is an unordered model that passes the arithmetic mean of the input word embeddings through a multilayer perceptron and feed the final layer’s representation to a softmax layer. We choose DAN because it ignores cross-lingual variations in word orders (i.e., syntax) and thus generalizes well in CLDC. Despite its simplicity, DAN has state-of-the-art accuracies on both monolingual (Iyyer et al., 2015) and cross-lingual document classification (Chen et al., 2018).
Let $v_1, v_2, \cdots, v_n$ be the input word embeddings. A DAN uses the average of the word embeddings as the document representation $z_0$:

$$z_0 = \text{mean}(v_1, v_2, \cdots, v_n),$$  \hspace{1cm} (4)

and $z_0$ is passed through $k$ layers of non-linearity:

$$z_i = g(W_i \cdot z_{i-1} + b_i),$$  \hspace{1cm} (5)

where $i$ ranges from 1 to $k$, and $g$ is a non-linear activation function. The final representation $z_k$ is passed to a softmax layer to obtain a distribution over the label $y$.

$$p(y \mid w) = \text{softmax}(W_{k+1}z_k + b_{k+1}).$$  \hspace{1cm} (6)

**2.2 Character-Level Cross-Lingual Transfer**

To transfer character-level information across languages, the embedder uses the same character embeddings for both languages. The character-level BI-LSTM vocabulary is the union of the alphabets for the two languages, and the embedder does not differentiate identical characters from different languages. For example, a Spanish “a” has the same character embedding as a French “a”. Consequently, the embedder maps words with similar forms from both languages to similar vectors.

If the source language and the target language are lexically similar, the embedder can generalize knowledge learned about source language words to target language words through shared orthographic features. As an example, if the model learns that the Spanish word “religioso” (religious) is predictive of label $y$, the model automatically infers that “religioso” in Italian is also predictive of $y$, even though the model never sees any Italian text.

**2.3 Training Objective**

Our main objective is supervised document classification. We jointly train the classifier and the embedder to minimize average negative log-likelihood on labeled source language documents $S$:

$$L_s(\theta) = -\frac{1}{|S|} \sum_{(w,y)} \log p(y \mid w),$$  \hspace{1cm} (7)

where $\theta$ is model parameter, and $S$ contains source language examples with words $w$ and label $y$.

Sometimes we have additional resources for the source or target language. We use them to improve our model with multi-task learning (Collobert et al., 2011) via three auxiliary tasks.

**Word Translation (DICT).** There are many patterns when translating cognate words between related languages. For example, Italian “e” often becomes “ie” when translating into Spanish. “Tempo” (time) in Italian becomes “tiempo” in Spanish, and “concerto” (concert) in Italian becomes “concierto”
in Spanish. The embedder can learn these word translation patterns from a bilingual dictionary.

Let $\mathcal{D}$ be a bilingual dictionary with a set of word pairs $\langle w_s, w_t \rangle$, where $w_s$ and $w_t$ are translations of each other. We add a term to our objective to minimize average squared Euclidean distances between the embeddings of translation pairs (Mikolov et al., 2013):

$$L_d(\theta) = \frac{1}{|\mathcal{D}|} \sum_{\langle w_s, w_t \rangle} \| e(w_s) - e(w_t) \|^2_2.$$  

(8)

**Mimicking Word Embeddings (MIM).** Monolingual text classifiers often benefit from initializing embeddings with word vectors pre-trained on large unlabeled corpus (Collobert et al., 2011). This semi-supervised learning strategy helps the model generalize to word types outside labeled training data. Similarly, our embedder can MIMICK (Pinter et al., 2017) an existing source language word embeddings to learn and transfer useful representations.

Suppose we have a pre-trained source language word embedding matrix $\mathbf{X}$ with $V$ rows. The $i$-th row $\mathbf{x}_i$ is a vector for the $i$-th word type $w_i$. We add an objective to minimize the average squared Euclidean distances between the output of the embedder and $\mathbf{X}$:

$$L_e(\theta) = \frac{1}{V} \sum_{i=1}^V \| e(w_i) - \mathbf{x}_i \|^2_2.$$  

(9)

**Knowledge Distillation.** Sometimes we have a reliable reference classifier in another high-resource language $\mathcal{H}$ (e.g., English). If we have parallel text between $\mathcal{S}$ and $\mathcal{H}$, we can use knowledge distillation (Xu and Yang, 2017) to supply additional training signal. Let $\mathcal{P}$ be a set of parallel documents $\langle \mathbf{w}_s, \mathbf{w}_h \rangle$, where $\mathbf{w}_s$ is from source language $\mathcal{S}$, and $\mathbf{w}_h$ is the translation of $\mathbf{w}$ in $\mathcal{H}$. We add another objective to minimize the average Kullback-Leibler divergence between the predictions of our model and the reference model:

$$L_p(\theta) = \frac{1}{|\mathcal{P}|} \sum_{\langle \mathbf{w}_s, \mathbf{w}_h \rangle} \text{KL}(p_h(y | \mathbf{w}_h) \| p(y | \mathbf{w}_s)), $$  

(10)

where $p_h$ is the output of the reference classifier (in language $\mathcal{H}$), and $p$ is the output of CACO. In $\S$ 4, we mark models that use knowledge distillation with a superscript “P”.

We train on the four tasks jointly. Our final objective is:

$$L(\theta) = L_d(\theta) + \lambda_d L_d(\theta) + \lambda_e L_e(\theta) + \lambda_p L_p(\theta),$$  

(11)

where the hyperparameters $\lambda_d$, $\lambda_e$, and $\lambda_p$ trade off between the four tasks.

### 3 Related Work

Previous CLDC methods are typically word-based and rely on one of the following cross-lingual signals to transfer knowledge: large bilingual lexicons (Shi et al., 2010; Andrade et al., 2015), MT systems (Banea et al., 2008; Wan, 2009; Zhou et al., 2016), or cross-lingual word representations (Klementiev et al., 2012; Chen et al., 2018). Unfortunately, these resources are not available in every language. Recent work proposes unsupervised training methods for CLWE (Zhang et al., 2017a,b; Conneau et al., 2018; Artetxe et al., 2018a; Alvarez-Melis and Jaakkola, 2018; Hoshen and Wolf, 2018) and MT (Artetxe et al., 2018b; Lample et al., 2018a,b) without using any cross-lingual signal. However, these methods still require large monolingual corpora in the target language, and they might fail when the monolingual corpora for the two languages come from different domains (Søgaard et al., 2018). In contrast, CACO is much more data-efficient. By exploiting character-level similarities between related languages, CACO trained with few or no target language data is competitive with CLWE-based models.

Our work builds on the success of character-level Bi-LSTM embedders in monolingual NLP tasks, including language modeling and part-of-speech tagging (Ling et al., 2015), named entity recognition (Lample et al., 2016), and dependency parsing (Ballesteros et al., 2015). Character-level embedders can generate useful representations for rare and unseen words, especially for morphologically rich languages. We extend this to CLDC. Recent work also uses character Bi-LSTMs to directly reconstruct pre-trained monolingual word embeddings (Pinter et al., 2017), which motivates our MIMICK objective.

Cross-lingual transfer at the character-level is successfully used in low-resource paradigm completion (Kann et al., 2017), morphological tagging (Cotterell and Heigold, 2017), POS tagging (Kim et al., 2017), and named entity recognition (Cotterell and Duh, 2017; Lin et al., 2018), where the authors train a character-level model
Table 1: Comparison of models used in our experiments (introduced in §4.2). We compare CACO variants with two high-resource models: a CLWE-based model (CLWE) and a lightly supervised target language model (SUP). We also experiment with a model that combines CLWE with CACO (COM).

| Source labeled data | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
|---------------------|---|---|---|---|---|---|---|---|
| Target labeled data |   |   | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| CLWE feature        |   |   |   |   | ✓ | ✓ |   |   |
| Character feature   | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Small dictionary    | ✓ | ✓ |   |   |   |   |   |   |
| Source embedding    | ✓ | ✓ | ✓ |   |   |   |   |   |
| Knowledge distillation | ✓ | ✓ |   |   |   |   |   |   |

4 Experiments

When the source language and the target language are related, we expect that character-level knowledge transfer is more data-efficient than word-level knowledge transfer. We test this by comparing CACO models trained in low-resource settings and with CLWE-based models trained in high-resource settings on two CLDC datasets. We also compare CACO with a supervised monolingual model. On two datasets, CACO models are competitive with the baselines despite training on much less target language data. Finally, we train models that combine CACO with CLWE, which have significantly higher accuracy than models with only CLWE as features. This confirms that subword similarities between related languages effectively transfer knowledge for CLDC.

4.1 Classification Dataset

Our first dataset is Reuters multilingual corpus (RCV2), a collection of news stories labeled with topics (Lewis et al., 2004). Following Klementiev et al. (2012), we remove documents with multiple topic labels. For each language, we sample 1,500 training documents and 200 test documents with balanced labels. We conduct CLDC experiments between two North Germanic languages, Danish (DA) and Swedish (SV), and three Romance languages, French (FR), Italian (IT), and Spanish (ES).

To test CACO on truly low-resource languages, we build a second CLDC dataset with famine-related documents sampled from Tigrinya (TI) and Amharic (AM) LORELEI language packs (Strassel and Tracey, 2016). We train models to classify whether the document describes widespread crime or not. The Amharic language pack does not have annotations, so we manually label Amharic sentences based on English reference translations. Our final dataset contains 394 Tigrinya and 370 Amharic documents with balanced labels.

4.2 Models

We compare CACO trained low-resource settings with word-based models that use more resources. Table 1 summarizes our models.

CACO Variants. We experiment with several variants of CACO that uses different resources. The SRC model is only trained on labeled source language documents and do not use any unlabeled data. The DICT model is trained on both labeled documents and parallel dictionary with the supervised objective and the word translation objective. The MIM model is jointly trained to label source language documents and mimick the source language part of the pre-trained CLWE. The ALL model is trained using both the word translation and the mimick tasks. We sometimes use knowledge dis-

---

3 Corporate/Industrial (CCAT), Economics (ECAT), Government/Social (GCAT), and Markets (MCAT).
| source | target | SRC | DICT | MIM | ALL | CLWE | SUP | COM |
|--------|--------|-----|------|-----|-----|------|-----|-----|
| DA SV  | 56.0   | 62.8| 60.4 | 62.9| 69.3| 59.7 | 69.7|
| SV DA  | 56.7   | 60.2| 58.4 | 62.2| 51.4| 40.8 | 67.5|
| FR ES  | 49.6   | 59.3| 48.3 | 57.4| 63.9| 56.6 | 70.8|
| IT ES  | 50.2   | 54.6| 51.4 | 54.7| 43.4| 56.6 | 63.5|
| ES FR  | 48.5   | 49.7| 49.2 | 48.8| 63.1| 48.9 | 61.3|
| IT FR  | 45.9   | 52.1| 46.6 | 48.2| 26.7| 48.9 | 62.8|
| FR IT  | 43.3   | 53.2| 44.3 | 51.2| 43.6| 44.9 | 60.2|
| ES IT  | 49.7   | 53.5| 53.4 | 52.5| 51.3| 44.9 | 59.7|
| average| 50.0   | 55.7| 51.5 | 54.7| 51.6| 51.9 | 64.5|

Table 2: CLDC experiments between eight related European language pairs on RCV2 topic identification. The CACO models are competitive with word-based models that use more resources. The combined model (COM) has the highest average test accuracy. We **boldface** the best result for each row.

| source | target | SRC | MIM  | SRC* | MIM* | CLWE | COM |
|--------|--------|-----|------|------|------|------|-----|
| AM TI  | 55.5   | 56.3| 57.0 | 57.6 | 59.1 | 60.1 |
| TI AM  | 56.8   | 55.1| *    | *    | 58.1 | 59.5 |

Table 3: CLDC experiments between Amharic and Tigrinya on LORELEI disaster response dataset. CACO models are only slightly worse than CLWE-based models without using any target language data. For AM-TI, knowledge distillation (SRC* and MIM*) further improves CACO models. We do not experiment with knowledge distillation on TI because we cannot find enough unlabeled parallel text in the language pack. Combining CACO with pre-trained CLWE gives the highest test accuracy.

tillation to provide more classification signals for some models. We mark these models with a superscript “P”.

**CLWE-Based Model.** For comparison, we train word-based DANs with pre-trained CLWE features. We use the multiCCA embeddings (Ammar et al., 2016), which are trained on large corpora with millions of tokens and high-coverage dictionaries with hundreds of thousands of word types. The CLWE-based DANs are strong high-resource models that require large corpora from similar domains and dictionary for the target language. In contrast, we train CACO models in a simulated low-resource setting with few or no target language data. Therefore, it is unfair to directly compare the results of CACO and CLWE-based models. However, CACO models are often competitive with CLWE-based models in our experiment, demonstrating the effectiveness of character-level transfer learning.

**Supervised Model.** We also compare CACO with a lightly-supervised monolingual model (SUP), a word-based DAN trained on fifty labeled target language documents. Without using any target language supervision, CACO models has similar (and sometimes higher) test accuracies as SUP, showing that cross-lingual subword similarity is a valuable classification signal.

**Combined Model.** Finally, we experiment with a model that combines CACO and CLWE (COM) by feeding pre-trained CLWE as additional features for the classifier of a CACO model (SRC variant). The combined model on average has much higher accuracy than both CACO variants and CLWE-based model, showing that character-level knowledge transfer is useful even when we have enough unlabeled data to train high-quality CLWE.

### 4.3 Dictionary and Parallel Text

Some of the CACO models use a dictionary to learn word translation patterns. We train them on the same training dictionary used for pre-training the CLWE. To simulate the low-resource setting, we
| source | target | SRC | DICT | MIM | ALL | CLWE |
|--------|--------|-----|------|-----|-----|------|
| DA     | ES     | 32.5| 34.8 | 30.6| 38.2| **65.7** |
| DA     | FR     | 34.1| 41.8 | 35.5| 43.3| **45.9** |
| DA     | IT     | 36.8| 43.7 | 37.2| 41.5| **47.4** |
| SV     | ES     | 35.2| 42.5 | 34.6| 46.8| **48.5** |
| SV     | FR     | 27.4| 29.9 | 29.1| 28.3| **49.0** |
| SV     | IT     | 34.6| 36.4 | 33.3| 35.2| **40.4** |
| average|        | 33.4| 38.2 | 33.4| 37.2| **49.5** |

(a) North Germanic to Romance

| source | target | SRC | DICT | MIM | ALL | CLWE |
|--------|--------|-----|------|-----|-----|------|
| ES     | DA     | 47.7| 48.3 | 46.1| 52.0| **56.7** |
| ES     | SV     | 50.6| **53.7**| 48.5| 51.4| 52.4 |
| FR     | DA     | 46.7| 44.2 | 44.7| **48.6**| 45.3 |
| FR     | SV     | 52.9| 53.2 | 53.6| 52.8| **57.2** |
| IT     | DA     | 36.6| 43.6 | 34.8| 43.0| **48.2** |
| IT     | SV     | 37.8| 45.3 | 30.7| 43.9| **31.1** |
| average|        | 45.4| 48.1 | 43.1| **48.6**| 48.5 |

(b) Romance to North Germanic

Table 4: CLDC experiments between languages from different families on RCV2. When transferring from a North Germanic language to a Romance language, CACO models score much lower than CLWE-based models (left). Surprisingly, CACO models are on par with CLWE-based when transferring from a Romance language to a North Germanic language (right). We **boldface** the best result for each row.

Sample only 100 translation pairs from the original dictionary for CACO.7

The Amharic labeled dataset is very small compared to other languages,8 so we experiment with knowledge distillation for the Amharic to Tigrinya CLDC experiment using English-Amharic parallel text. We first train a reference English DAN on a large collection of labeled English documents compiled from other LORELEI language packs. We then use the knowledge distillation objective to train the CACO models to match the output of the English model on 1,200 English-Amharic parallel documents sampled from the Amharic language pack.9 We do not use knowledge distillation on other language pairs, because we have enough labeled examples for the RCV2 languages, and we do not have enough unlabeled parallel text in the Tigrinya language pack.

4.4 Effectiveness of CACO

We train each model using ten different random seeds and report their average test accuracy.10 We describe training details and hyperparameters in Appendix. Table 2 and 3 show average test accuracies of different models on RCV2 and LORELEI on nine related language pairs.

---

7Pilot experiments confirm that a larger dictionary can help, but we focus on the low-resource setting here.

8Each Amharic document only contains one sentence.

9To avoid introducing extra label bias, we sample the parallel documents such that the English model output approximately follows a uniform distribution.

10For models that use dictionaries, we also re-sample the training dictionary for each run.

Character-Level Knowledge Transfer. The low-resource character-based CACO models have similar average test accuracy as the high-resource word-based models. The SRC variant does not use any target language data, and yet it has higher accuracy than CLWE-based models on three language pairs. The accuracies are also similar to a word-based model supervisedly trained with fifty labeled target language documents (SUP). When we already have a good CLWE, we can get the best of both worlds by combining them (COM), which has the highest average test accuracy. These results show that character-level knowledge transfer is sample-efficient and complementary to word-level knowledge transfer.

Multi-Task Learning. Training CACO with multi-task learning further improves the accuracy. For almost all language pairs, the multi-task CACO variants have higher test accuracies than SRC. On RCV2, word translation (DICT) is particularly effective, even with only 100 translation pairs. Interestingly, word translation and mimick tasks together (ALL) do not consistently increase the accuracy over only using the dictionary (DICT). On the LORELEI dataset where labeled document is limited, knowledge distillation (SRCP and MIMP) also significantly increases accuracies.

Language Relatedness. We expect character-level knowledge transfer to be less effective on language pairs when the source language and the target language are less close to each other. For comparison, we experiment on RCV2 with transferring between more distantly related language
pairs: a North Germanic language and a Romance language (Table 4). Indeed, CACO models score consistently lower than the CLWE-based models when transferring from a North Germanic source language to a Romance target language. However, CACO models are surprisingly competitive with CLWE-based models when transferring from the opposite direction. This asymmetry is likely due to morphological differences between the two language families. Unfortunately, our datasets only have a limited number of language families. We leave a more systematic study on how language proximity affect the effectiveness of CACO to future work.

**Multi-Source Transfer.** On RCV2, we experiment with training CACO models on two Romance languages and testing on a third Romance language. Languages can be similar along different dimensions, and therefore adding more source languages may be beneficial. Moreover, using multiple source languages has a regularization effect and prevents the model from overfitting to a single source language. For fair comparison, we sample 750 training documents from each source language, so that the multi-source models are still trained on 1,500 training documents (like the single-source models). We use a similar strategy to sample the training dictionaries and pre-trained word embeddings. Multi-source models (Table 5) consistently have higher accuracies than single-source models (Table 2).

**Learned Word Representation.** We evaluate the word representations learned by CACO with a word translation task between French, Italian, and Spanish. Specifically, we use the SRC embedder to generate embeddings for all French, Italian, and Spanish words that appear in multiCCA’s vocabulary. Table 6 shows the top-1 word translation accuracy on the test dictionaries from MUSE (Conneau et al., 2018).11 Although the SRC embedder is not exposed to any cross-lingual signal, it still rivals CLWE on the word translation task by exploiting subword similarities between languages.

**Qualitative Analysis.** To understand how cross-lingual subword similarity can help document classification, we manually compare the output of a CLWE-based model and a CACO model (DICT variant) from the Spanish to Italian CLDC experiment. We inspect their predictions on some Italian documents. Sometimes CACO avoids the mistakes of CLWE-based models by correctly aligning word pairs that are misaligned in the pre-trained CLWE. For example, in the CLWE, “relevancia” (relevance) is the closest Spanish word for the Italian word “interesse” (interest), while the CACO embedder maps both the Italian word “interesse” (interest) and the Spanish word “interesse” (interest) to the same point. Consequently, CACO correctly classifies an Italian document about the interest rate with GCAT (government), while the CLWE-based model predicts MCAT (market).

| source | target | SRC | DICT | MIM | ALL |
|--------|--------|-----|------|-----|-----|
| FR/IT  | ES     | 58.8| 67.0 | 55.8| 65.3|
| ES/IT  | FR     | 51.8| 55.8 | 50.3| 56.0|
| ES/FR  | IT     | 53.2| 56.1 | 55.9| 56.5|
|        | average| 54.6| 59.6 | 54.0| 59.3|

Table 5: Results of CLDC experiments using two source languages. Models trained on two source languages are generally better than models trained on only one source language (Table 2).

| source | target | CLWE | CACO |
|--------|--------|------|------|
| ES     | FR     | 36.8 | 31.1 |
| ES     | IT     | 44.0 | 33.1 |
| FR     | ES     | 34.0 | 30.9 |
| FR     | IT     | 33.5 | 29.6 |
| IT     | ES     | 42.1 | 37.5 |
| IT     | FR     | 35.6 | 36.4 |
|        | average| 37.7 | 33.1|

Table 6: Word translation accuracies (P@1) for different embeddings. The CACO embeddings are generated by the embedder of a SRC model trained on the source language. Without any cross-lingual signal, the CACO embedder has competitive word translation accuracy as CLWE pre-trained on large target language corpora and dictionaries.

5 Conclusion and Future Work

We investigate character-level knowledge transfer between related languages for CLDC. Our transfer learning scheme, CACO, exploits subword similarities between related languages through shared character representations to generalize from source language data. We provide empirical evaluation of CACO on multiple related language pairs and show that character-level knowledge transfer between related language is highly effective.

---

11 Each word is translated with nearest-neighbor search.
References

David Alvarez-Melis and Tommi S. Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In Proceedings of Empirical Methods in Natural Language Processing.

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

Daniel Andrade, Kunihiko Sadamasa, Akihiro Tamura, and Masaaki Tsuchida. 2015. Cross-lingual text classification using topic-dependent word probabilities. In Conference of the North American Chapter of the Association for Computational Linguistics.

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

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In Proceedings of the International Conference on Learning Representations.

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with LSTMs. In Empirical Methods in Natural Language Processing.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In Proceedings of Empirical Methods in Natural Language Processing.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. Transactions of the Association for Computational Linguistics, 6:557–570.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12:2493–2537.

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

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In International Joint Conference on Natural Language Processing.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In Empirical Methods in Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Conference of the North American Chapter of the Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framework for phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6):602–610.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In Proceedings of Empirical Methods in Natural Language Processing.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In Proceedings of the Association for Computational Linguistics.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. One-shot neural cross-lingual transfer for paradigm completion. In Proceedings of the Association for Computational Linguistics.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In Proceedings of Empirical Methods in Natural Language Processing.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In International Conference on Learning Representations.

Alexandre Klementiev, Ivan Titov, and Binod Bhattacharjee. 2012. Inducing crosslingual distributed representations of words. Proceedings of International Conference on Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In Conference of the North American Chapter of the Association for Computational Linguistics.

Guillaume Lample, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In Proceedings of the International Conference on Learning Representations.

Guillaume Lample, Mylène Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In Proceedings of Empirical Methods in Natural Language Processing.
David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5(Apr):361–397.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the Association for Computational Linguistics*.

Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fernandez, Silvio Amir, Luis Marujo, and Tiago Luís. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of Empirical Methods in Natural Language Processing*.

David R. Mortensen, Siddharth Dalmia, and Patrick Litell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Language Resources and Evaluation Conference*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword rns. In *Proceedings of Empirical Methods in Natural Language Processing*.

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

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings 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 Language Resources and Evaluation Conference*.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Association for Computational Linguistics*.

Michelle Yuan, Benjamin Van Durme, and Jordan Boyd-Graber. 2018. Multilingual anchoring: Interactive topic modeling and alignment across languages. In *Proceedings of Advances in Neural Information Processing Systems*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017a. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the Association for Computational Linguistics*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017b. Earth mover’s distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of 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 Association for Computational Linguistics*.
A Training Details

For the CLWE-based models, we use forty dimensional multiCCA word embeddings (Ammar et al., 2016). We use three ReLU layers with 100 hidden units and 0.1 dropout for the CLWE-based DAN models and the DAN classifier of the CACO models. The BI-LSTM embedder uses ten dimensional character embeddings and forty hidden states with no dropout. The outputs of the embedder are forty dimensional word embeddings. We set $\lambda_d$ to 1, $\lambda_e$ to 0.001, and $\lambda_p$ to 1 in the multi-task objective (Equation 11). The hyperparameters are tuned in a pilot Italian-Spanish CLDC experiment using held-out datasets.

All models are trained with Adam (Kingma and Ba, 2015) with default settings. We run the optimizer for a hundred epochs with mini-batches of sixteen documents. For models that use additional resources, we also sample sixteen examples from each type of training data (translation pairs, pre-trained embeddings, or parallel text) to estimate the gradients of the auxiliary task objectives $L_d$, $L_e$, and $L_p$ (defined in §2.3) at each iteration.