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

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

Yoshinari Fujinuma
Computer Science
University of Colorado
Boulder, CO, USA
fujinumay@gmail.com

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

Abstract

Text classification must sometimes be applied in a low-resource language with no labeled training data. 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 (**?**). Despite this revolution for high-resource languages such as English, some languages are left behind because of the dearth of text data generally and specifically labeled data. Often, the need for a text classifier in a low-resource language is acute, as text classifiers can provide situational awareness in emergent incidents (**?**). Cross-lingual document classification (**?**, 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 (**?**; **?**), machine translation (**?**; **?**; **?**, MT), topic models (**?**; **?**), cross-lingual word embeddings (**?**, CLWE), or multilingual contextualized embeddings (**?**) 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. These similarities are often not found at word-level but at character-level. Therefore, we investigate 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, **C**lassification **A**ided by **C**onvergent **O**rthography (CACO) that capitalizes on character-level similarities between related language pairs. Previous CLDC methods treat words as atomic symbols and do not transfer character-level 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 then uses the *shared* representation across languages to label the document. The embedder learns morpho-semantic regularities, while the classifier connects lexical semantics to labels.

To allow cross-lingual transfer, we use a *single* model with shared character embeddings for both 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 can be fairly accurate 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 model with multi-task learning. We encourage the embedder to produce similar word embeddings for translation pairs from a dictionary, which captures patterns between cognates. We also teach the embedder to MIMICK pre-trained word embeddings in the source lan-

*Now at Google Research Zürich

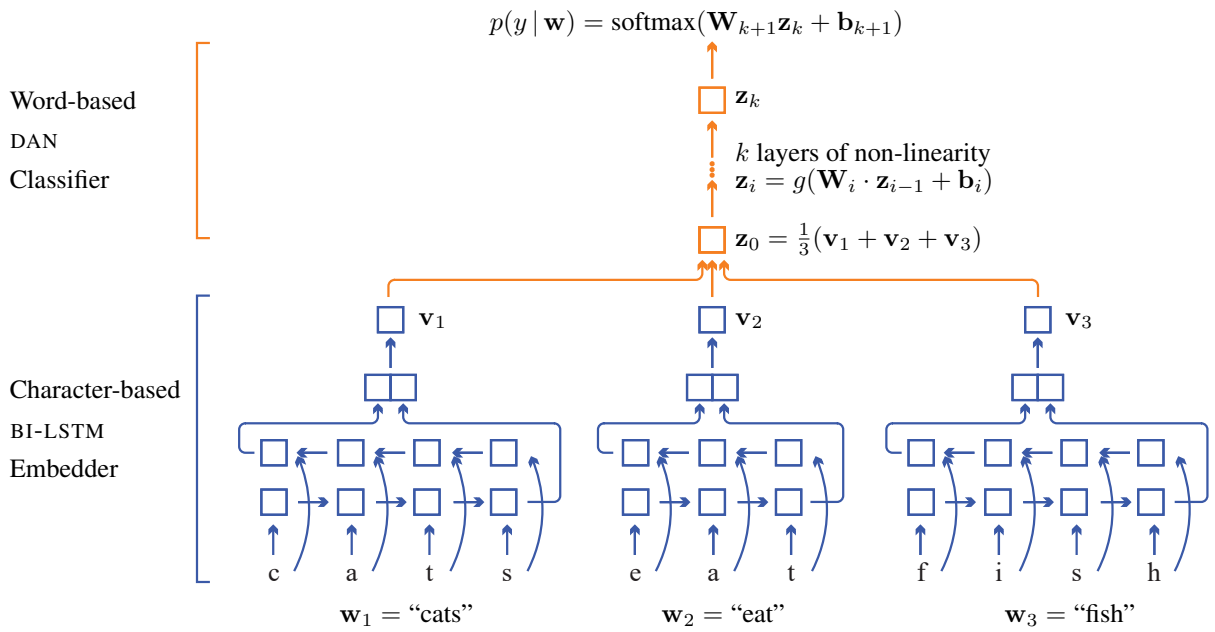

Figure 1: Computation graph of CACO on an example sentence ("cats eat fish"). *Bottom:* Each input word $\mathbf{w}_i$ is mapped to a vector $\mathbf{v}_i$ by passing its characters through a BI-LSTM embedder. *Top:* Word vectors $\{\mathbf{v_i}\}$ are then passed through a DAN classifier to predict the label y. Specifically, DAN transforms the average of the word vectors $\mathbf{z}_0$ with k layers of non-linearity and a final softmax layer.

guage (**?**), which exposes the model to more word types. When we have a good reference model in another high-resource language, we can train our model to make similar predictions as the reference model on parallel text (**?**).

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 also useful when we have enough resources to train good CLWE—using CLWE as extra features, CACO outperforms the baseline CLWE-based models by a large margin.

2 CACO: Classification Aided by Convergent Orthography

This section introduces our method, CACO, which trains a multilingual document classifier using labeled datasets in a source language $\mathcal{S}$ and applies the classifier to a low-resource target language $\mathcal{T}$. We focus on the setting where $\mathcal{S}$ and $\mathcal{T}$ are related and have similar orthographic features.

2.1 Model Architecture

Let $\mathbf{x}$ be an input document with a sequence of words $\mathbf{x} = \langle \mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_n \rangle$, where each word $\mathbf{w}_i$ is a sequence of character. Our model maps the document $\mathbf{x}$ to a distribution over possible labels y in two steps (Figure 1). First, we generate a word embedding $\mathbf{v}_i$ for each input word $\mathbf{w}_i$ using a character-based embedder e:

$$\mathbf{v}_i = e(\mathbf{w}_i). \tag{1}$$

We then feed the word embeddings to a word-based classifier f to compute the distribution over labels y:

$$p(y \mid \mathbf{w}) = f\left(\langle \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n \rangle\right). \tag{2}$$

We can use any sequence model for the embedder e and the classifier f. For our experiments, we use a bidirectional LSTM (**?**, BI-LSTM) embedder and a deep averaging network (**?**, DAN) classifier.

BI-LSTM Embedder. BI-LSTM is a powerful sequence model that captures complex non-local dependencies. Character-based BI-LSTM embedders are used in many natural language processing tasks (**?**; **?**; **?**). To embed a word $\mathbf{w}$, we pass the character sequence $\mathbf{w}$ to a left-to-right LSTM and the reversed character sequence $\mathbf{w'}$ to a right-to-left LSTM. We concatenate the final hidden states of the two LSTM and apply a linear transformation:

$$e(\mathbf{w}) = \mathbf{W}_e \cdot [\overrightarrow{\text{LSTM}}(\mathbf{w}); \overleftarrow{\text{LSTM}}(\mathbf{w'})] + \mathbf{b}_e, \tag{3}$$

where the functions $\overrightarrow{\text{LSTM}}$ and $\overleftarrow{\text{LSTM}}$ 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 feeds the final layer's representation to a softmax layer. DAN ignores cross-lingual variations in word order (i.e., syntax) and thus generalizes well in

CLDC. Despite its simplicity, DAN has near state-of-the-art accuracies on both monolingual (**?**) and cross-lingual document classification (**?**).

Let $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ be the word embeddings generated by the character-based embedder. DAN uses the average of the word embeddings as the document representation $\mathbf{z}_0$:

$$\mathbf{z}_0 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}_i, \tag{4}$$

and $\mathbf{z}_0$ is passed through k layers of non-linearity:

$$\mathbf{z}_i = g(\mathbf{W}_i \cdot \mathbf{z}_{i-1} + \mathbf{b}_i), \tag{5}$$

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

$$p(y \mid \mathbf{x}) = \text{softmax}(\mathbf{W}_{k+1}\mathbf{z}_k + \mathbf{b}_{k+1}). \tag{6}$$

We use the same classifier parameters $\mathbf{W}_i$ across languages. In other words, the DAN classifier is language-independent. This is possible because the embedder generates consistent word representations across related languages, which we discuss in the next section.

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 orthographically 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.

In our experiments, we focus on related language pairs that share the same script. For related languages with different scripts, we can apply CACO to the output of a transliteration tool or a grapheme-to-phoneme transducer (**?**). We leave this to future work.

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:

$$L_s(\boldsymbol{\theta}) = -\frac{1}{|S|} \sum_{\langle \mathbf{x}, y \rangle} \log p(y \mid \mathbf{x}), \tag{7}$$

where $\boldsymbol{\theta}$ is a vector representing all model parameters, and S is a set of source language examples with words $\mathbf{x}$ and label y.

Sometimes we have additional resources for the source or target language. We use them to improve CACO with multi-task learning (**?**) 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" in 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 D be a bilingual dictionary with a set of word pairs $\langle \mathbf{w}_s, \mathbf{w}_t \rangle$, where $\mathbf{w}_s$ and $\mathbf{w}_t$ are translations of each other. We add a term to our objective to minimize average squared Euclidean distances between the embeddings of translation pairs (**?**):

$$L_d(\boldsymbol{\theta}) = \frac{1}{|D|} \sum_{\langle \mathbf{w}_s, \mathbf{w}_t \rangle} \|e(\mathbf{w}_s) - e(\mathbf{w}_t)\|_2^2. \tag{8}$$

Mimicking Word Embeddings (MIM). Monolingual text classifiers often benefit from initializing embeddings with word vectors pre-trained on large unlabeled corpus (**?**). This semi-supervised learning strategy helps the model generalize to word types outside labeled training data. Similarly, our embedder can MIMICK (**?**) an existing *source language* word embeddings to generalize better.

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

$$L_e(\boldsymbol{\theta}) = \frac{1}{V} \sum_{i=1}^{V} \|e(\mathbf{w}_i) - \mathbf{E}_i\|_2^2. \tag{9}$$

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 (**?**) to supply additional training signal. Let P be a set of parallel documents $\langle \mathbf{x}_s, \mathbf{x}_h \rangle$, where $\mathbf{x}_s$ is from source language $\mathcal{S}$, and $\mathbf{x}_h$ is the translation of $\mathbf{x}_s$ in $\mathcal{H}$. We add another objective term to minimize the average Kullback-Leibler divergence between the predictions of our model and the reference model:

$$L_p(\boldsymbol{\theta}) = \frac{1}{|P|} \sum_{\langle \mathbf{x}_s, \mathbf{x}_h \rangle \in P} \text{KL}(p_h(y \mid \mathbf{x}_h) \parallel p(y \mid \mathbf{x}_s)), \tag{10}$$

where p_h is the output of the reference classifier (in language $\mathcal{H}$), and p is the output of CACO. In § 3, we mark models that use knowledge distillation with a superscript "P".

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

$$L(\boldsymbol{\theta}) = L_s(\boldsymbol{\theta}) + \lambda_d L_d(\boldsymbol{\theta}) + \lambda_e L_e(\boldsymbol{\theta}) + \lambda_p L_p(\boldsymbol{\theta}), \tag{11}$$

where the hyperparameters λ_d, λ_e, and λ_p trade off between the four tasks.

	CACO				Baseline		
	SRC	DICT	MIM	ALL	CLWE	SUP	COM
Source labeled data	✓	✓	✓	✓	✓	✓	✓
Pre-trained source embedding			✓	✓			
Small dictionary		✓		✓			
Pre-trained CLWE					✓		✓
Target labeled data						✓	
RCV2 average accuracy	50.0	55.7	51.5	54.7	51.6	51.9	**64.5**

Table 1: Comparison of models used in our experiments (introduced in Section 3.2). For each model, we list its required resources and average accuracy on RCV2 over eight related language pairs (accuracy for each pair in Table 2). We compare CACO variants with two high-resource models: a CLWE-based model (CLWE) and a lightly supervised target language model (SUP). Both baselines require more target language resources than CACO variants, and yet they have lower average accuracy than some CACO variants, which confirms that character-level knowledge transfer is highly efficient. We also experiment with a model that combines CLWE with CACO (COM). This combined model has the highest average accuracy, indicating that CLWE and CACO are complementary when both options are available.

3 Experiments

When the source language and the target language are related, we expect character-level knowledge transfer to be more data-efficient than word-level knowledge transfer because character-level transfer allows generalization across words with similar forms. 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 both datasets, CACO models have similar average accuracy as the baselines *while requiring 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. These results confirms that character-level similarities between related languages effectively transfer knowledge for CLDC.

3.1 Classification Dataset

Our first dataset is Reuters multilingual corpus (RCV2), a collection of news stories labeled with four topics (**?**):Corporate/Industrial (CCAT), Economics (ECAT), Government/Social (GCAT), and Markets (MCAT). Following **?** (**?**), 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 (**?**). We train binary classifiers to detect whether the document describes widespread crime or not. For Tigrinya documents, the labels are extracted from the situation frame annotation in the language pack. We mark all documents with a "widespread crime/violence" situation frame as positive. The Amharic language pack does not have annotations, so we label Amharic sentences based on English reference translations included from the language pack. Our dataset contains 394 Tigrinya and 370 Amharic documents with balanced labels.

3.2 Models

We compare CACO trained under 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 uses the least amount of resource. It is only trained on labeled source language documents and do not use any unlabeled data. The **DICT** model requires a dictionary and is trained with the word translation auxiliary task. The **MIM** model requires a pre-trained source language embedding and uses the mimick auxliliary task. The **ALL** model is the most expensive variant. It is trained with both the word translation and the mimick auxiliary tasks. In LORELEI experiments, we also use knowledge distillation to provide more classification signals for some models. We mark these models with a superscript "P".

CLWE-Based Model. Our first word-based model is a DAN with pre-trained multiCCA CLWE features (**?**). The CLWE are trained on large target language corpora with millions of tokens and high-coverage dictionaries with hundreds of thousands of word types. In contrast, we train CACO models in a simulated low-resource setting with few or no target language data. Despite the resource gap, CACO models have similar average test accuracy as CLWE-based models, demonstrating the effectiveness of character-level transfer learning.

Supervised Model. Next, we compare CACO with a lightly-supervised monolingual model (**SUP**), a word-based DAN trained on fifty labeled target language documents. We only apply this baseline to RCV2, because the labeled document sets in LORELEI are too small to split further. The su-

source	target	CACO				CLWE	SUP	COM
		SRC	DICT	MIM	ALL			
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 average accuracy of CACO models are competitive with word-based models that use *more resources* such as target language corpora or labeled data (Table 1). The combined model (COM) has the highest average test accuracy. We **boldface** the best result for each row.

source	target	CACO				CLWE	COM
		SRC	MIM	SRC^P	MIM^P		
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^P and MIM^P) 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.

pervised model requires labeled target language documents, which often do not exist in labeled documents. Without using any target language supervision, CACO models have similar (and sometimes higher) test accuracies as SUP, showing that CACO effectively learns from a related language.

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). This model requires the same amount of resource as the CLWE-based model. 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.

3.3 Auxiliary Task Data

Some of the CACO models (DICT and ALL) 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 sample **only 100 translation pairs** from the original dictionary for CACO. Pilot experiments confirm that a larger dictionary can help, but we focus on the low-resource setting where only a small dictionary is available.

The Amharic labeled dataset is very small compared to other languages because each Amharic example only contains one sentence. As introduced in Section 2.3, one way to provide additional training signal is by knowledge distillation from a third high-resource language. For the Amharic to Tigrinya CLDC experiment, we apply knowledge distillation 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. To avoid introducing extra label bias, we sample the parallel documents such that the English model output approximately follows a uniform distribution.

We do not use knowledge distillation on other language pairs. For RCV2, we already have enough labeled examples and therefore do not need knowledge distillation. For Tigrinya to Amharic CLDC experiment, we do not have enough unlabeled parallel text in the Tigrinya language pack to apply knowledge distillation.

3.4 Training Details

For CLWE-based models, we use forty dimensional multi-CCA word embeddings (**?**). 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 λ_d to 1, λ_e to 0.001, and λ_p to 1 in the multi-task objec-

source	target	CACO SRC	CACO DICT	CACO MIM	CACO 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	CACO SRC	CACO DICT	CACO MIM	CACO 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.

tive (Equation 11). The hyperparameters are tuned in a pilot Italian-Spanish CLDC experiment using held-out datasets.

All models are trained with Adam (**?**) 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 Section 2.3) at each iteration.

3.5 Effectiveness of CACO

We train each model using ten different random seeds and report their average test accuracy. For models that use dictionaries, we also re-sample the training dictionary for each run. Table 1 compares resource requirement and average RCV2 accuracy of CACO and baselines. Table 2 and 3 show test accuracies on nine related language pairs from RCV2 and LORELEI.

Character-Level Knowledge Transfer. Experiments confirm that character-level knowledge transfer is sample-efficient and complementary to word-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 its average test accuracy on RCV2 (50.0%) is very close to the CLWE model (51.6%) and the supervised model SUP (51.6%). When we already have a good CLWE, we can get the best of both worlds by combining them (COM), which has a much higher average test accuracy (64.5%) than CACO and the two baselines.

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. It increases average test accuracy from 50.0% to 55.7%, outperforming both word-based baseline models. 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 (SRC^{P} and MIM^{P}) also increases accuracies by around 1.5%.

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. Languages can be similar along different dimensions, and therefore adding more source languages may be beneficial. On RCV2, we experiment with training CACO models on *two* Romance languages and testing on a third Romance language. 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. Word translation is a popular intrinsic evaluation task for cross-lingual word representations. Therefore, we evaluate the word representations learned by the BI-LSTM embedder on a word translation benchmark. Specifically, we use the SRC embedder to generate embeddings for all French, Italian, and Spanish words that appear in multiCCA's vocabulary and translate each word with nearest-neighbor search. Table 6 shows the top-1 word translation accuracy on the test dictionaries from MUSE (**?**). Although the SRC embedder is not exposed to any cross-lingual signal, it rivals CLWE on the word translation task by exploiting character-level similarities between languages.

Qualitative Analysis. To understand how cross-lingual character-level similarity helps classification, we manually compare the output of a CLWE-based model and a CACO model (DICT variant) from the Spanish to Italian CLDC experiment. 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).

4 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 (**?**; **?**), MT systems (**?**; **?**; **?**), or CLWE (**?**). One exception is the recently proposed multilingual BERT (**?**; **?**), which uses a subword vocabulary. Unfortunately, some languages do not have these resources. CACO can help bridge the resource gap. By exploiting character-level similarities between related languages, CACO can work effectively with few or no target language data.

To adapt CLWE to low-resource settings, recent unsupervised CLWE methods (**?**; **?**) do not use dictionary or parallel text. These methods can be further improved with careful normalization (**?**) and interactive refinement (**?**). However, unsupervised CLWE methods still require large monolingual corpora in the target language, and they might fail when the monolingual corpora of the two languages come from different domains (**?**; **?**) and when the two language have different morphology (**?**). In contrast, CACO does not require any target language data.

Cross-lingual transfer at character-level is successfully used in low-resource paradigm completion (**?**), morphological tagging (**?**), part-of-speech tagging (**?**), and named entity recognition (**?**; **?**; **?**; **?**), where the authors train a character-level model jointly on a small labeled corpus in target language and a large labeled corpus in source language. Our method is similar in spirit, but we focus on CLDC, where it is less obvious if orthographic features are helpful. Moreover, we introduce a novel multi-task objective to use different types of monolingual and cross-lingual resources.

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

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

Acknowledgement

We thank the members of UMD CLIP and the anonymous reviewers for their feedback. Zhang and Boyd-Graber are supported by DARPA award HR0011-15-C-0113 under subcontract to Raytheon BBN Technologies. Fujinuma and Boyd-Graber are supported by NSF grant IIS-1564275. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

Quae iure magni sit neque tempora dignissimos doloribus aliquid asperiores quod deleniti, nobis dicta ullam eligendi accusantium quaerat at autem, aliquid odit numquam eum nemo architecto dignissimos.Corrupti eum illo similique libero voluptate debitis facilis earum nesciunt eius omnis, iste earum repellendus velit impedit quasi quos, beatae nostrum quis dignissimos quam atque quo mollitia provident veniam, rem tenetur quidem iusto quibusdam nostrum odio labore modi ut, sed voluptates consequuntur cum similique excepturi voluptatibus qui ea et error.Cumque quasi commodi, quae sequi est nulla illo soluta ratione ut reiciendis vel,