		CACO				
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

		CACO				
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.

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$^{\text{P}}$ and MIM$^{\text{P}}$) 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,