A Little Annotation does a Lot of Good:
A Study in Bootstrapping Low-resource Named Entity Recognizers

Aditi Chaudhary,   Jiateng Xie,   Zaid Sheikh,
Graham Neubig,   Jaime G. Carbonell
{*aschaudh, jiatengx, zsheikh, gneubig, jgc*}@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University

Abstract

Most state-of-the-art models for named entity recognition (NER) rely on the availability of large amounts of labeled data, making them challenging to extend to new, lower-resourced languages. However, there are now several proposed approaches involving either cross-lingual transfer learning, which learns from other highly resourced languages, or active learning, which efficiently selects effective training data based on model predictions. This paper poses the question: given this recent progress, and limited human annotation, what is the most effective method for efficiently creating high-quality entity recognizers in under-resourced languages? Based on extensive experimentation using both simulated and real human annotation, we find a dual-strategy approach best, starting with a cross-lingual transferred model, then performing targeted annotation of only uncertain entity spans in the target language, minimizing annotator effort. Results demonstrate that cross-lingual transfer is a powerful tool when very little data can be annotated, but an entity-targeted annotation strategy can achieve competitive accuracy quickly, with just one-tenth of training data. The code is publicly available here.[1]

1 Introduction

Named entity recognition (NER) is the task of detecting and classifying named entities in text into a fixed set of pre-defined categories (person, location, etc.) with several downstream applications including machine reading (Chen et al., 2017), entity and event co-reference (Yang and Mitchell, 2016), and text mining (Han and Sun, 2012). Recent advances in deep learning have yielded state-of-the-art performance on many sequence labeling tasks, including NER (Collobert et al., 2011;

---

[1] https://github.com/Aditi138/EntityTargetedActiveLearning

Ma and Hovy, 2016; Lample et al., 2016; Peters et al., 2018). However, the performance of these models is highly dependent on the availability of large amounts of annotated data, and as a result their accuracy is significantly lower on languages that have fewer resources than English. In this work, we ask the question "how can we efficiently bootstrap a high-quality named entity recognizer for a low-resource language with only a small amount of human effort?" Specifically, we leverage recent advances in data-efficient learning for low-resource languages, proposing the following "recipe" for bootstrapping low-resource entity recognizers: First, we use *cross-lingual transfer learning* (Yarowsky et al., 2001; Ammar et al., 2016), which applies a model trained on another language to low-resource languages, to provide a good preliminary model to start the bootstrapping process. Specifically, we use the model of Xie et al. (2018), which reports strong results on a number of language pairs. Next, on top of this transferred model we further employ *active learning* (Settles and Craven, 2008; Marcheggiani and Artieres, 2014), which helps improve annotation efficiency by using model predictions to select informative, rather than random, data for human annotators. Finally, the model is *fine-tuned* on data obtained using active learning to improve accuracy in the target language.

Within this recipe, the choice of specific method for choosing and annotating data within active learning is highly important to minimize human effort. One relatively standard method used in previous work on NER is to select full sequences based on a criterion for the uncertainty of the entities recognized therein (Culotta and McCallum, 2005). However, as it is often the case that only a single entity within the sentence may be of interest, it can still be tedious and wasteful to annotate full sequences when only a small portion of the
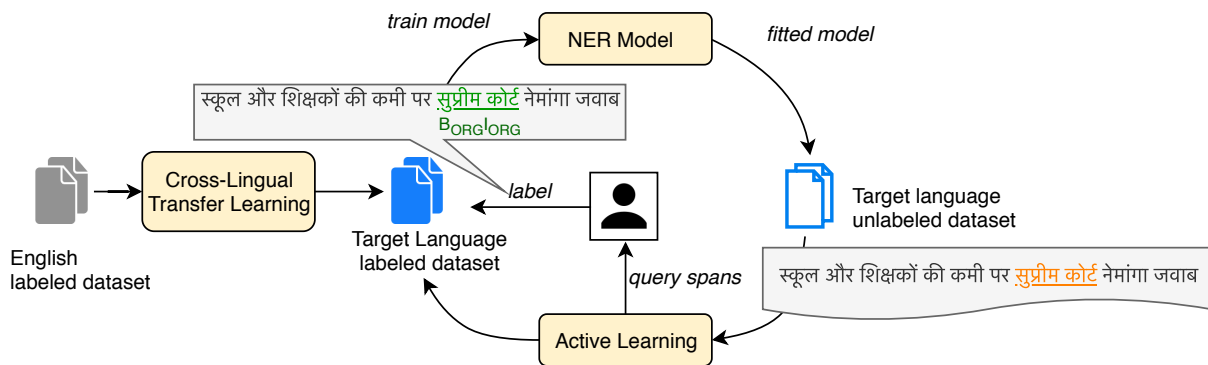
Figure 1: Our proposed recipe: cross-lingual transfer is used for projecting annotations from an English labeled dataset to the target language. Entity-targeted active learning is then used to select informative sub-spans which are likely entities for humans to annotate. Finally, the NER model is fine-tuned on this partially-labeled dataset.

sentence is of interest (Neubig et al., 2011; Sperber et al., 2014). Inspired by this finding and considering the fact that named entities are both important and sparse, we propose an entity-targeted strategy to save annotator effort. Specifically, we select uncertain subspans of tokens within a sequence that are most likely named entities. This way, the annotators only need to assign types to the chosen subspans without having to read and annotate the full sequence. To cope with the resulting partial annotation of sequences, we apply a constrained version of conditional random fields (CRFs), partial CRFs, during training that only learn from the annotated subspans (Tsuboi et al., 2008; Wanvarie et al., 2011).

To evaluate our proposed methods, we conducted simulated active learning experiments on 5 languages: Spanish, Dutch, German, Hindi and Indonesian. Additionally, to study our method in a more practical setting, we conduct human annotation experiments on two low-resource languages, Indonesian and Hindi, and one simulated low-resource language, Spanish. In sum, this paper makes the following contributions:

1. We present a bootstrapping recipe for improving low-resource NER. With just one-tenth of tokens annotated, our proposed entity-targeted active learning method provides the best results among all active learning baselines, with an average improvement of 9.9 F1.

2. Through simulated experiments, we show that cross-lingual transfer is a powerful tool, outperforming the un-transferred systems by an average of 8.6 F1 with only one-tenth of tokens annotated.

3. Human annotation experiments show that annotators are more accurate in annotating entities when using the entity-targeted strategy as opposed to full sequence annotation. Moreover, this strategy minimizes annotator effort by requiring them to label fewer tokens than the full-sequence annotation.

2 Approach

As noted in the introduction, our bootstrapping recipe consists of three components (1) cross-lingual transfer learning, (2) active learning to select relevant parts of the data to annotate, and (3) fine-tuning of the model on these annotated segments. Steps (2) and (3) are continued until the model has achieved an acceptable level of accuracy, or until we have exhausted our annotation budget. The system overview can be seen in Figure 1. In the following sections, we describe each of these three steps in detail.

2.1 Cross-lingual Transfer Learning

The goal of cross-lingual learning is to take a recognizer trained in a source language, and transfer it to a target language. Our approach to doing so for NER follows that of Xie et al. (2018), and we provide a brief review in this section.

To begin with, we assume access to two sets of pre-trained monolingual word embeddings in the source and target languages, X and Y, one small bilingual lexicon, either provided or obtained in an unsupervised manner (Artetxe et al., 2017; Conneau et al., 2017a), and labeled training data in the source language. Using these resources, we train bilingual word embeddings (BWE) to create a word-to-word translation dictionary, and finally

use this dictionary to translate the source training data into the target language, which we use to train an NER model.

To learn BWE, we first obtain a linear mapping W by solving the following objective:

$$W^* = \underset{W}{\arg\min} \|WX_D - Y_D\|_F \text{ s.t. } WW^\top = I,$$

where X_D and Y_D correspond to the aligned word embeddings from the bilingual lexicon. F denotes the Frobenius norm. We can first compute the singular value decomposition $Y_D^T X_D = U \sum V^\top$, and solve the objective by taking $W^* = UV^\top$. We obtain BWE by linearly transforming the source and target monolingual word embeddings with U and V, namely XU and YV.

After obtaining the BWE, we find the nearest neighbor target word for every source word in the BWE space using the cross-domain similarity local scaling (CSLS) metric (Conneau et al., 2017b), which produces a word-to-word translation dictionary. We use this dictionary to translate the source training data into the target language, and simply copy the label for each word, which yields transferred training data in the target language. We train an NER model on this transferred data as our preliminary model. Going forward, we refer to the use of cross-lingual transferred data as CT.

2.2 Entity-Targeted Active Learning

After training a model using cross-lingual transfer learning, we start the active learning process based on this model's outputs. We begin by training a NER model Θ using the above model's outputs as training data. Using this trained model, our proposed entity-targeted active learning strategy, referred as ETAL, then selects the most informative spans from a corpus D of unlabeled sequences. Given an unlabeled sequence s, ETAL first selects a span of tokens $s_i^j = s_i \cdots s_j$ such that s_i^j is a likely named entity, where $i, j \in [0, |s|]$. Then, in order to obtain highly informative spans across D, ETAL computes the entropy H for each occurrence of the span s_i^j in D and then aggregates them over the entire corpus D, given by:

$$H_{aggregate}(s_i^j) = \sum_{x_i^j \in D} H(x_i^j) \mathbb{1}(x_i^j = s_i^j)$$

where x is an unlabeled sequence in D. Finally, the spans s_i^j with the highest aggregate uncertainty $H_{aggregate}$ are selected for manual annotation.

We now describe the procedure for calculating $H(x_i^j)$, which is the entropy of a span x_i^j being a likely entity. Given an unlabeled sequence x, the trained NER model Θ is used for computing the marginal probabilities $p_\theta(y_i|\mathbf{x})$ for each token x_i across all possible labels $y_i \in Y$ using the forward-backward algorithm (Rabiner, 1989), where Y is the set of all labels. Using these marginals we calculate the entropy of a given span x_i^j being an entity as shown in Algorithm 1.

Algorithm 1 Entity-Targeted Active Learning

1:   $B \leftarrow$ *label-set denoting beginning of an entity*
2:   $I \leftarrow$ *label-set denoting inside of an entity*
3:   $O \leftarrow$ *outside of an entity span*
4:   $p_\theta(y_i|\mathbf{x}) \leftarrow$ *marginal probability of label y_i*
5:   *for token x_i*
6:
7:   **for** $i \leftarrow 1...len(x), j = 1$ **do**
8:     $p_{span}^{ij} = \sum_{y \in B} p_\theta(y_i|\mathbf{x})$
9:     **for** $j \leftarrow i + 1...len(x)$ **do**
10:       $p_{entity} = p_{span}^{ij} * p_\theta(O_j|\mathbf{x})$
11:       $H = entropy\,(p_{entity})$
12:       **if** $H > H_{threshold}$ **then**
13:         $H_{aggregate}(x_i^j) + = H$
14:       $p_{span}^{ij} = p_{span}^{ij} * \sum_{y \in I} p_\theta(y_j|\mathbf{x})$

Let B denote the set of labels indicating beginning of an entity, I the set of labels indicating inside of an entity and O denoting outside of an entity. First, we compute the probability of a span x_i^j being an entity, starting with the token i, by marginalizing $p_\theta(y_i|\mathbf{x})$ over all labels in B, denoted as p_{span}^{ij}. Since an entity can span multiple tokens, for each subsequent token j being part of that entity, we marginalize $p_\theta(y_j|\mathbf{x})$ over all labels in I and combine it with p_{span}^{ij}. Finally, we compute $p_{entity} = p_{span}^{ij} * p_\theta(O_j|\mathbf{x})$, which denotes end of a likely entity. Since we use the marginal probability for computing p_{entity}, it already factors in the transition probability between tags. Thus, any invalid sequences such as $B_{PER}I_{ORG}$ have low scores. Since contiguous spans have overlapping tokens, using dynamic programming (DP) to compute p_{span}^{ij} avoids an exponential computation when considering all possible spans in a sequence. Using p_{entity}, we compute the entropy H and only consider the spans having H higher than a pre-defined threshold $H_{threshold}$. The reason for this thresholding is purely for computational purposes as it allows us to discard all spans

that have a very low probability of being an entity, keeping the number of spans actually stored in memory low. As mentioned above, we aggregate the entropy of spans $H_{aggregate}$ over the entire unlabeled set, thus combining uncertainty sampling with a bias towards high frequency entities.

Using this strategy, we select subspans in each sequence for annotation. The annotator only needs to assign named entity types to the chosen subspans, adjust the span boundary if needed, and ignore the rest of the sequence, saving much effort.

2.3 Training the NER model

With the newly obtained training data from active learning, we attempt to improve the original transferred model. In this section, we first describe our model architecture, and try to address: 1) how to train the NER model effectively with partially annotated sequences? 2) what training scheme is best suited to improve the transferred model?

2.3.1 Model Architecture

Our NER model is a BiLSTM-CNN-CRF model based on Ma and Hovy (2016) consisting of: a character-level CNN, that allows the model to capture subword information; a word-level BiLSTM, that consumes word embeddings and produces context sensitive hidden representations; and a linear-chain CRF layer that models the dependency between labels for inference. We use the above model for training the initial NER model on the transferred data as well as for re-training the model on the data acquired from active learning.

2.3.2 PARTIAL-CRF

Active learning with span-based strategies such as ETAL, produces a training dataset of partially labeled sequences. To train the NER model on these partially labeled sequences, we take inspiration from Bellare and McCallum (2007); Tsuboi et al. (2008) and use a constrained CRF decoder. Normally, CRF computes the likelihood of a label sequence $\mathbf{y}$ given a sequence $\mathbf{x}$ as follows:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{\prod_{t=1}^T \psi_i(y_{t-1}, y_t, \mathbf{x}, t)}{Z(\mathbf{x})}$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y} \in \mathbf{Y}(T)} \prod_{t=1}^T \psi_i(y_{t-1}, y_t, \mathbf{x}, t)$$

where T is the length of the sequence, $\mathbf{Y}(T)$ denotes the set of all possible label sequences with length T, and $\psi_i(y_{t-1}, y_t, \mathbf{x}) = \exp(\mathbf{W}_{y_{t-1}, y_t}^T \mathbf{x}_i + \mathbf{b}_{y_{t-1}, y_t})$ is the energy function. To compute the likelihood of a sequence where some labels are unknown, we use a constrained CRF which marginalizes out the un-annotated tokens. Specifically, let $\mathbf{Y_L}$ denote the set of all possible sequences that include the partial annotations (for un-annotated tokens, all labels are possible), and we compute the likelihood as: $p_\theta(\mathbf{Y_L}|\mathbf{x}) = \sum_{\mathbf{y} \in \mathbf{Y_L}} p_\theta(\mathbf{y}|\mathbf{x})$, referred as PARTIAL-CRF.

2.3.3 Training Scheme

To improve our model with the newly labeled data, we directly fine-tune the initial model, trained on the transferred data, on the data acquired through active learning, referred as FINETUNE. Each active learning run produces more labeled data, for which this training procedure is repeated again. We also compare the NER performance using two other training schemes: CORPUSAUG, where we train the model on the concatenated corpus of transferred data and the newly acquired data, and CORPUSAUG+FINETUNE, where we additionally fine-tune the model trained using CORPUSAUG on just the newly acquired data.

3 Experiments

In this section, we evaluate the effectiveness of our proposed strategy in both simulated (§3.2) and human-annotation experiments (§3.3).

3.1 Experimental Settings

Datasets: The first evaluation set includes the benchmark CoNLL 2002 and 2003 NER datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) for Spanish (from the Romance family), Dutch and German (like English, from the Germanic family). We use the standard corpus splits for train/dev/test. The second evaluation set is for the low-resource setting where we use the Indonesian (from the Austronesian family), Hindi (from the Indo-Aryan family) and Spanish datasets released by the Linguistic Data Consortium (LDC).[2] We generate the train/dev/test split by random sampling. Details of the corpus statistics are in the Appendix.

English-transferred Data: We use the same experimental settings and resources as described in Xie et al. (2018) to get the translations of the English training data for each target language.

---

[2]LDC2017E62,LDC2016E97,LDC2017E66

Active Learning Setup: As described in Section §2.2, a DP-based algorithm is employed to select the uncertain entity spans which runs for all n-grams having length $<= 5$. This length was approximated by computing the 90th percentile on the length of entities in the English training data. $H_{threshold}$ is a hyper-parameter set to 1e-8. The details of the NER model hyper-parameters can be found in the Appendix.

3.2 Simulation Experiments

Setup: We use cross-lingual transfer (§2.1) to train our initial NER model and test on the target language. This is the same setting as Xie et al. (2018) and serves as our baseline. Then we use several active learning strategies to select data for manual annotation using this trained NER model. We compare our proposed ETAL strategy with the following baseline strategies:

SAL: Select whole sequences for which the model has least confidence in the most likely labeling (Culotta and McCallum, 2005).

CFEAL: Select least confident spans within a sequence using the confidence field estimation method (Culotta and McCallum, 2004).

RAND: Select spans randomly from the unlabeled set for annotation.

In this experimental setting, we simulate manual annotation by using gold labels for the data selected by active learning. At each subsequent run, we annotate 200 tokens and fine-tune the NER model on all the data acquired so far, which is then used to select data for the next run of annotation.

3.2.1 Results

Figure 2 summarizes the results for all datasets across the different experimental settings. Each data-point on the x-axis corresponds to the NER performance after annotating 200 additional tokens. CT denotes using cross-lingual transferred data to train the initial NER model for both kick-starting the active learning process and also for fine-tuning the NER model on the newly-acquired data. PARTIAL-CRF/FULL-CRF denote the type of CRF decoder used in the NER model. Throughout this paper, we report results averaged across all active learning runs unless otherwise noted. Individual scores are reported in the Appendix.

As can be seen in the figure, our proposed recipe ETAL+PARTIAL-CRF+CT outperforms the previous

Dataset	Tokens	ETAL	SAL	RAND	CFEAL
Hindi LDC	200	54.8 ± 2.6	49.6 ± 2.8	50.2 ± 0.4	50.4 ± 2.8
	600	64.7 ± 2.6	51.5 ± 2.9	56.1 ± 2.6	54.3 ± 2.5
	1200	69.9 ± 2.5	56.6 ± 2.7	56.8 ± 2.6	64.4 ± 2.7
Indonesian LDC	200	**47.4± 2.6**	**47.9 ± 2.4**	**46.8 ± 2.3**	**48.5 ± 2.3**
	600	54.5 ± 2.4	44.5 ± 2.2	47.2 ± 2.2	46.0 ± 2.3
	1200	60.5 ± 2.3	44.7 ± 2.3	51.9 ± 2.3	49.5 ± 2.3
Spanish LDC	200	66.3 ± 3.8	62.0 ± 3.6	61.2 ± 1.2	62.5 ± 3.7
	600	65.8 ± 4.1	62.5 ± 3.7	61.9 ± 2.0	63.8 ± 3.7
	1200	78.9 ± 3.5	62.3 ± 3.6	64.6 ± 4.0	68.6 ± 3.9

Table 1: Variance analysis for significance testing of different active learning systems using paired bootstrap resampling. ± denotes the 95% confidence intervals. Systems which are not statistically significant than the best system ETAL are in bold. The CoNLL datasets reflect the same observation, as can be seen in Appendix.

active learning baselines for all the datasets. Holding the other two components of CT and PARTIAL-CRF constant, we conduct experiments to compare the different active learning strategies, which are denoted by the solid lines in Figure 2. We see that ETAL outperforms the other strategies by a significant margin for both the CoNLL datasets: German (+6.1 F1), Spanish (+5.3 F1), Dutch (+6.3 F1) and the LDC datasets: Hindi (+9.3 F1), Indonesian (+9.0 F1), Spanish (+7.5 F1), at the end of all runs. Furthermore, even with just one-tenth annotated tokens, the proposed recipe is only (avg.) -5.2 F1 behind the model trained using all labeled data, denoted by SUPERVISED ALL. Although CFEAL also selects informative spans, ETAL outperforms it because ETAL is optimized to select likely entities, causing more entities to be annotated for Hindi (+43), Indonesian (+207), Spanish-CoNLL (+1579), German (+906), Dutch (+836), except for Spanish-LDC (-184). Despite fully labeled data being adding in SAL, ETAL outperforms it because SAL selects longer sentences with fewer entities: Hindi (-934), Indonesian (-1290), Spanish-LDC (-527), Spanish-CoNLL (-2395), German (-2086), Dutch (-2213).

From Figure 2 we see that ETAL performs better than the baselines across multiple runs. To verify that this is not an artifact of randomness in the test data, we use a paired bootstrap resampling method, as illustrated in Koehn (2004), to compare SAL, CFEAL, RAND with ETAL. For each system, we compute the F1 score on randomly sampled 50% of the data and perform 10k bootstrapping steps at three active learning runs. From Table 1 we see that the baselines are significantly worse than ETAL at 600 and 1200 annotated tokens.
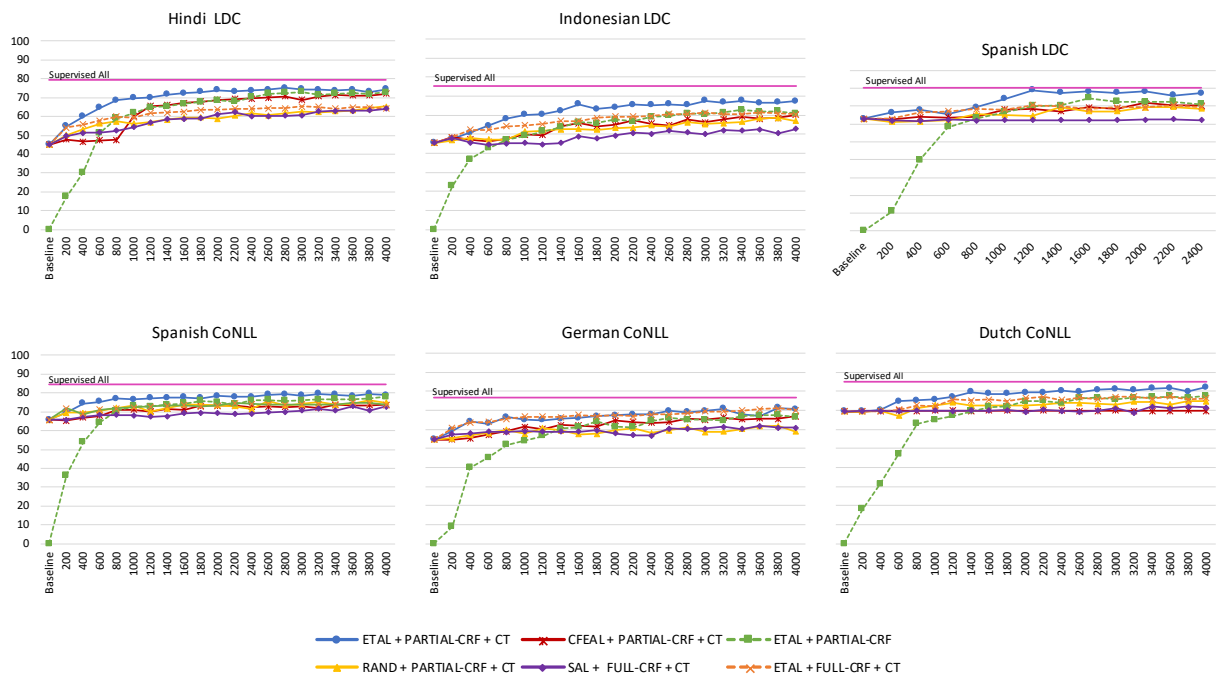
Figure 2: Comparison of the NER performance trained with the FineTune scheme, across six datasets. Solid lines compare the different active learning strategies. Dashed lines show the ablation experiments. The x-axis denotes the total number of tokens annotated and the y-axis denotes the F1 score.

3.2.2 Ablation Study

In order to study the contribution of CT and PARTIAL-CRF in improving the NER performance, we conduct the following ablation, denoted by dashed lines in Figure 2.

CT: We observe that the transferred data from English provides a good start to the NER model: 69.4 (Dutch), 63.0 (Spanish-LDC), 65.7 (Spanish-CoNLL), 54.7 (German), 45.4 (Indonesian), 45.0 (Hindi) F1. As expected, cross-lingual transfer helps more for the languages closely related to English which are Dutch, German, Spanish. For this ablation, we train a ETAL+PARTIAL-CRF where no transferred data is used. Therefore, to create the seed data, we randomly annotate 200 tokens in the target language and thereafter use ETAL. We observe that as more in-domain data is acquired, the un-transferred setting soon approaches the transferred setting ETAL+PARTIAL-CRF+CT suggesting that an efficient annotation strategy can help close the gap between these two systems with as few as ∼1000 tokens (avg.).

PARTIAL-CRF: We study the effect of using the original CRF (FULL-CRF) instead of the PARTIAL-CRF for training with partially labeled data. Since the former requires fully labeled sequences, the un-annotated tokens in a sequence are labeled with the model predictions. We see from Figure 2 that the ETAL+FULL-CRF+CT performs worse (avg. -4.1 F1) than ETAL+PARTIAL-CRF+CT. This is because the FULL-CRF significantly hurts the recall, as much as by an average of -11.0 points for Hindi, -1.4 for Indonesian, -7.4 for Spanish-LDC, -3.3 for German, -3.7 for Dutch, -4.8 for Spanish CoNLL.

3.2.3 Comparison of Training Schemes

We experiment with different NER training regimes (described in §2.3.3) for ETAL. We observe that generally fine-tuning not only speeds up the training but also gives better performance than CORPUSAUG. For brevity of space, we compare results for two languages in Figure 3:[3] Dutch, a relative of English, and Hindi, a distant language. We see that FINETUNE performs better for Hindi whereas CORPUSAUG+FINETUNE performs better for Dutch. This is because Dutch is closely related to English and benefits the most from the transferred data being explicitly augmented. Whereas for Hindi, which is typologically distant from English, the transferred data is noisy and thus the model doesn't gain much from the transferred data. Xie et al. (2018) make a similar observation in their

---

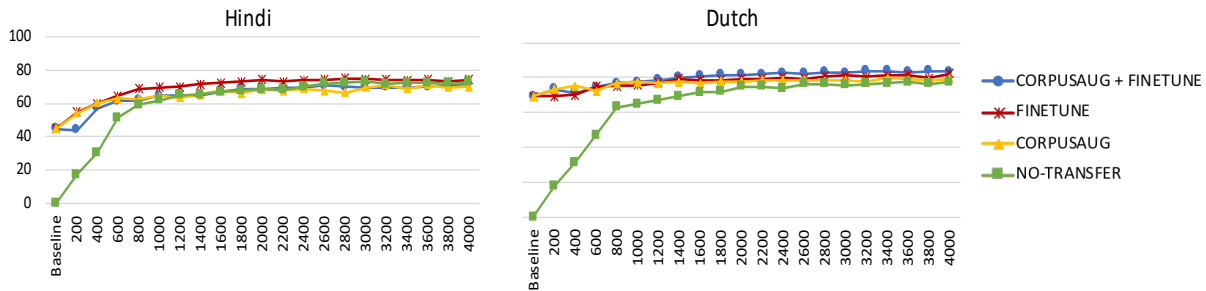[3]The results for other datasets can be found in Appendix.

5169

Figure 3: Comparison of the NER performance trained with different schemes for the ETAL strategy. The x-axis denotes the total number of tokens annotated and the y-axis denotes the F1 score.

experiments with German.

3.3 Human Annotation Experiments

Setup: We conduct human annotation experiments for Hindi, Indonesian and Spanish to understand whether ETAL helps reduce the annotation effort and improve annotation quality in practical settings. We compare ETAL with the full sequence strategy (SAL). We use six native speakers, two for each language, with different levels of familiarity with the NER task. Each annotator was provided with practice sessions to gain familiarity with the annotation guidelines and the user interface. The annotators annotated for 20 mins time for each strategy. For ETAL, the annotator was required to annotate single spans i.e each sequence contained one span of tokens. This involved assigning the correct label and adjusting the span boundary if required. For SAL, the annotator was required to annotate all possible entities in the sequence. We randomized the order in which the annotators had to annotate using ETAL and SAL strategy. Figure 5 illustrates the human annotation process for the ETAL strategy in the annotation interface. [4]

3.3.1 Results

Table 2 records the results of the human annotation experiments. We first compare each annotator's annotation quality with respect to the oracle under both ETAL and SAL, denoted by *Annotator Performance*. We see that both Hindi and Spanish annotators have higher annotation quality using ETAL. This is because by selecting possible entity spans, ETAL not only saves effort on searching the entities in a sequence but also allows the annotators to read less overall and concentrate more

---

[4]The code for the annotation interface can be found at https://gitlab.com/cmu_ariel/ariel-annotation-interface.

	Annotator Performance		Test Performance (# annotated tokens)		
	ETAL	SAL	ETAL	SAL	SAL-Full
HI-1	**78.8**	63.7	**50.4** (326)	44.2 (326)	53.3 (1894)
HI-2	**82.7**	72.2	**49.1** (234)	45.9 (234)	55.6 (2242)
ID-1	66.1	77.8	**50.4** (425)	45.8 (425)	51.3 (3232)
ID-2	73.0	79.5	**51.2** (251)	46.5 (251)	54.0 (2874)
ES-1	79.7	75.0	**63.7** (204)	62.2 (204)	64.6 (2134)
ES-2	83.1	70.4	**63.8** (199)	62.2 (199)	62.6 (2134)

Table 2: Annotator performance measures F1 of each annotator with respect to the oracle. Test Performance measures the NER F1 scores using the annotations as training data. The number in the brackets denote the number of annotated tokens used for training the NER model. ES:Spanish, HI:Hindi, ID: Indonesian.

on the things that they do read, as seen in Figure 4(a). However, for SAL we see that the annotator missed a likely entity because they focused on the other more salient entities in the sequence. For Indonesian, we see an opposite trend due to several inconsistencies in the gold labels. The most common inconsistency being when a common noun is part of an entity. For instance, the gold standard annotates the span *Kabupaten Bogor* as an entity where *Kabupaten* means "district". Whereas for ~~Kabupaten~~ *Aceh tengah*, the gold standard does not include *Kabupaten*. Similarly, the same span *gunung krakatau* is annotated inconsistently across different mentions where sometimes they exclude the *gunung* (mountain) token.Since the annotators referred to these examples during their practice session, their annotations had similar inconsistencies. This issue affects ETAL more than SAL because ETAL selects more entities for annotation.

The *Test Performance* compares the performance of the NER models trained with these annotations. The number in the brackets denotes the total number of annotated tokens used for train-
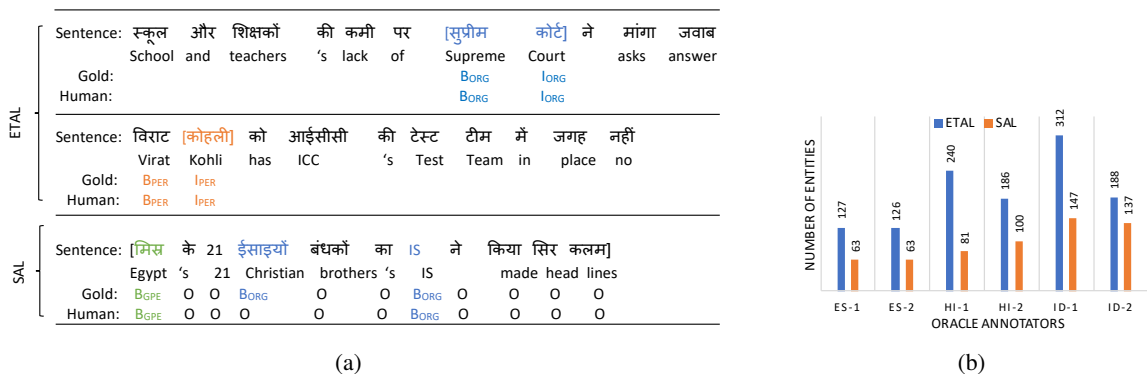
Figure 4: (a) Examples from Hindi human annotation experiments for both ETAL and SAL. Square brackets denote the spans (for ETAL) or the entire sequence (for SAL) selected by the respective active learning strategy. (b) Comparing the number of entities in the data selected by ETAL and SAL, as annotated by oracle.

ing the NER model. We observe that SAL has a larger number of annotated tokens than ETAL. This is because most sequences selected by SAL did not have any entities. Since "not-an-entity" is the default label in the annotation interface, no operation is required for annotating these, allowing for more tokens to be annotated per unit times. When we count the number of entities present in the data selected by the two strategies, we see in Figure 4(b) that data selected by ETAL has a significantly larger number of entities than SAL, across all the 6 annotation experiments. Therefore, we first compare the NER performance on the same number of annotated tokens. From Table 2 we see that under this setting ETAL outperforms SAL, similar to the simulation results. We note that when we consider all the annotated tokens, SAL (denoted by SAL-FULL) has slightly better results. However, despite having 6 times fewer annotated tokens, the difference between ETAL and SAL-FULL is (avg.) 2.1 F1. This suggests that ETAL can achieve competitive performance with fewer annotations.

From both the simulation and human experiments, we can conclude that a targeted annotation strategy such as ETAL achieves competitive performance with less manual effort while maintaining high annotation quality. Given that ETAL can help find twice as many entities as SAL, a potential application of ETAL can also be for creating a high-quality entity gazetteer under a short time budget. Since a naive strategy of SAL allows for more labelled data to be acquired in the same amount of time, in the future we plan to explore mixed-mode annotation where we choose either full sequences or spans for annotation.

4 Related Work

Cross-Lingual Transfer: Transferring knowledge from high-resource languages has been extensively used for improving low-resource NER. More common approaches rely on annotation projection methods where annotations in source language are projected to the target language using parallel corpora (Zitouni and Florian, 2008; Ehrmann et al., 2011) or bilingual dictionaries (Xie et al., 2018; Mayhew et al., 2017). Cross-lingual word embeddings (Bharadwaj et al., 2016; Chaudhary et al., 2018) also provide a way to leverage annotations from related languages.

Active Learning (AL): AL has been widely explored for many NLP tasks- NER: Shen et al. (2017) explore token-level annotation strategies, Chen et al. (2015) present a study on AL for clinical NER; Baldridge and Palmer (2009) evaluate how well AL works with annotator expertise and label suggestions, Garrette and Baldridge (2013) study type and token based strategies for low-resource languages. Settles and Craven (2008) present a nice survey on the different AL strategies for sequence labeling tasks, whereas Marcheggiani and Artieres (2014) discuss the strategies for acquiring partially labeled data. Wanvarie et al. (2011); Neubig et al. (2011); Sperber et al. (2014) show the advantages of training a model on this partially labeled data. All above methods focus on either token or full sequence annotation.
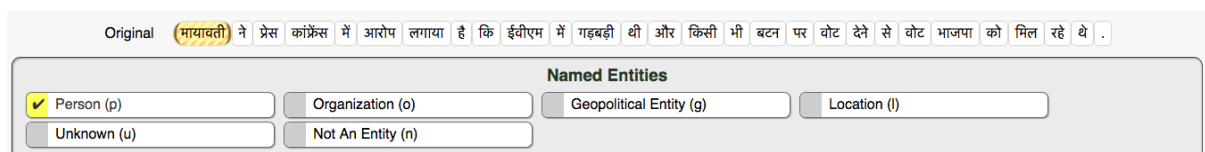
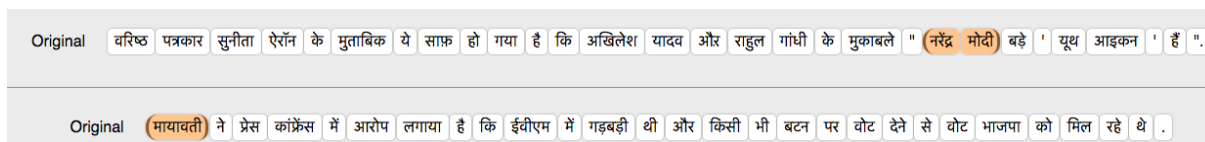The most similar work to ours perhaps is that of
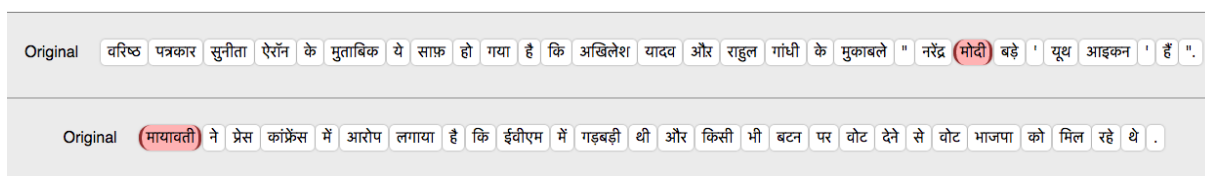
(a) Selected spans using ETAL strategy are highlighted for the human annotator to annotate.
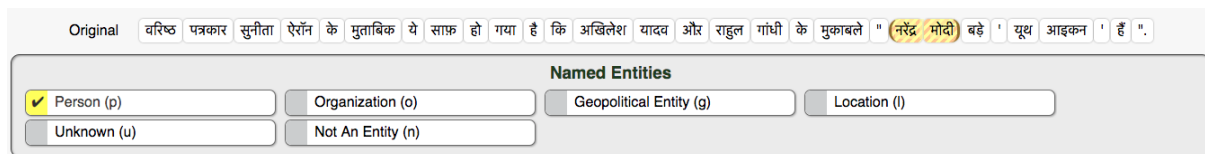
(b) Human annotator correcting the span boundary and assigning the correct entity type.

(c) Human annotator assigning the correct entity type only since selected span boundary is correct.

(d) Partially-annotated sequences after being annotated by the human annotator.

Figure 5: Example of the human annotation process for Hindi.

Fang and Cohn (2017), which selects informative word types for low-resource POS tagging. However, their method requires the annotator to annotate single tokens, which is not trivially applicable for multi-word entities in practical settings.

5   Conclusion

In this paper, we presented a study on how to efficiently bootstrap NER systems for low-resource languages using a combination of cross-lingual transfer learning and active learning. We conducted both simulated and human annotation experiments across different languages and found that: 1) cross-lingual transfer is a powerful tool, constantly beating systems without using transfer; 2) our proposed recipe works the best among known active learning baselines; 3) our proposed active learning strategy saves annotator much effort while ensuring high quality. In future, to account for different levels of annotator expertise, we plan to combine proactive learning (Li et al., 2017) with our proposed method.

Acknowledgement

References

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Kedar Bellare and Andrew McCallum. 2007. Learning extractors from unlabeled text using relevant databases. In *Sixth international workshop on information integration on the web*.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017a. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017b. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Aron Culotta and Andrew McCallum. 2004. Confidence estimation for information extraction. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 109–112. Association for Computational Linguistics.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 5, pages 746–751.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593, Vancouver, Canada. Association for Computational Linguistics.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Xianpei Han and Le Sun. 2012. An entity-topic model for entity linking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 105–115. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Maolin Li, Nhung Nguyen, and Sophia Ananiadou. 2017. Proactive learning for named entity recognition. In *BioNLP 2017*, pages 117–125, Vancouver, Canada,. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074. Association for Computational Linguistics.

Diego Marcheggiani and Thierry Artieres. 2014. An experimental comparison of active learning strategies for partially labeled sequences. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 898–906.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 529–533. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Lawrence R Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1070–1079. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Matthias Sperber, Mirjam Simantzik, Graham Neubig, Satoshi Nakamura, and Alex Waibel. 2014. Segmentation for efficient supervised language annotation with an explicit cost-utility tradeoff. *Transactions of the Association for Computational Linguistics*, 2:169–180.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Yuta Tsuboi, Hisashi Kashima, Hiroki Oda, Shinsuke Mori, and Yuji Matsumoto. 2008. Training conditional random fields using incomplete annotations. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 897–904. Association for Computational Linguistics.

Dittaya Wanvarie, Hiroya Takamura, and Manabu Okumura. 2011. Active learning with subsequence sampling strategy for sequence labeling tasks. *Information and Media Technologies*, 6(3):680–700.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299, San Diego, California. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 600–609. Association for Computational Linguistics.