Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities

Alexander Erdmann,[†‡] David Joseph Wrisley,[‡] Benjamin Allen,[†] Christopher Brown,[†]
Sophie Cohen-Bodénès,[*] Micha Elsner,[†] Yukun Feng,[†] Brian Joseph,[†]
Béatrice Joyeux-Prunel[*] and Marie-Catherine de Marneffe[†]
[†]Ohio State University, USA [‡]New York University Abu Dhabi, UAE
[*]Ecole Normale Supérieure, France
{ae1541,djw12}@nyu.edu,{allen.2021,brown.2583,elsner.14,
feng.749,joseph.1,demarneffe.1}@osu.edu,
sophiebodenes@gmail.com, beatrice.joyeux-prunel@ens.fr

Abstract

Scholars in inter-disciplinary fields like the Digital Humanities are increasingly interested in semantic annotation of specialized corpora. Yet, under-resourced languages, imperfect or noisily structured data, and user-specific classification tasks make it difficult to meet their needs using off-the-shelf models. Manual annotation of large corpora from scratch, meanwhile, can be prohibitively expensive. Thus, we propose an active learning solution for named entity recognition, attempting to maximize a custom model's improvement per additional unit of manual annotation. Our system robustly handles any domain or user-defined label set and requires no external resources, enabling quality named entity recognition for Humanities corpora where such resources are not available. Evaluating on typologically disparate languages and datasets, we reduce required annotation by 20-60% and greatly outperform a competitive active learning baseline.

1 Introduction

Reaping the benefits of recent advances in Named Entity Recognition (NER) is challenging when dealing with under-resourced languages, niche domains, imperfect or noisily structured data, or user-specific classification tasks. Scholars in inter-disciplinary fields like the Digital Humanities (DH) are increasingly interested in semantic annotation of specialized corpora that invoke many of these challenges. Thus, such corpora cannot easily be annotated automatically with blackbox, off-the-shelf NER models. Manual annotation of large corpora from scratch, meanwhile, can be prohibitively costly. Successful DH initiatives like the Pelagios Commons (Simon et al., 2016), which collects geospatial data from historical sources, often require extensive funding, relying on considerable manual annotation (Simon et al., 2017).

To this end, we introduce the Humanities Entity Recognizer (HER),[1] a whitebox toolkit for build-your-own NER models, freely available for public use. HER robustly handles any domain and user-defined label set, guiding users through an active learning process whereby sentences are chosen for manual annotation that are maximally informative to the model. Informativeness is determined based on novel interpretations of the *uncertainty*, *representativeness*, and *diversity* criteria proposed by Shen et al. (2004). In contrast to literature emphasizing the disproportionate or exclusive importance of uncertainty (Shen et al., 2017; Zhu et al., 2008; Olsson, 2009), we observe significant improvements by integrating all three criteria.

In addition to a robust active learning based NER toolkit, we also contribute a novel evaluation framework. This *inclusive* framework considers the accuracy with which an entire corpus is annotated, regardless of which instances are annotated manually versus automatically, such that no instance is held out when the active learning algorithm considers candidates for annotation. The standard, *exclusive* evaluation framework, by contrast, only measures the accuracy of the final trained model's predictions on a held out test set. Thus, the inclusive framework is relevant to the user who wants to annotate a finite corpus as fast and as accurately as possible by any means necessary, whereas the exclusive framework is relevant to the user who wants to build an NER tool that can generalize well to other corpora.

We conduct extensive experiments comparing several combinations of active learning algorithms and NER model architectures in both frameworks across many typologically diverse languages and domains. The systematic differences between inclusive and exclusive results demonstrate that while deep NER model architectures

[1] github.com/alexerdmann/HER.

Proceedings of NAACL-HLT 2019, pages 2223–2234
Minneapolis, Minnesota, June 2 - June 7, 2019.

(Lample et al., 2016) are highly preferable for tagging held out sentences, shallow models (Lafferty et al., 2001) perform better on sentences that could have been chosen for manual annotation but were not selected by the active learning algorithm. We argue for the importance of considering both frameworks when evaluating an active learning approach, as the intended application determines which framework is more relevant and thus, which model should be employed. Controlling for the NER model, HER's active learning sentence ranking component achieves significant improvement over a competitive baseline (Shen et al., 2017). Because HER does not reference the inference model during sentence ranking, this provides counter evidence to Lowell et al. (2018)'s hypothesis that *non-native* active learning is suboptimal.

2 Related Work

The best known NER systems among humanists are Stanford NER (Finkel et al., 2005), with pretrained models in several languages and an interface for building new models, and among researchers interested in NER for spatial research, the Edinburgh Geoparser (Grover et al., 2010), with fine grained NER for English. Erdmann et al. (2016) and Sprugnoli (2018), among others, have shown that such off-the-shelf models can be substantially improved on DH-relevant data. Work such as Smith and Crane (2001) and Simon et al. (2016) represent a large community mining such data for geospatial entities. Additional DH work on NER concerns the impact of input data structure and noisy optical character recognition (Van Hooland et al., 2013; Kettunen et al., 2017).

Low Resource NER Language agnostic NER is highly desirable, yet limited by the data available in the least resourced languages. Curran and Clark (2003) demonstrate that careful feature engineering can be typologically robust, though data hungry neural architectures have achieved state-of-the-art performance without feature engineering (Lample et al., 2016). To enable neural architectures in low resource environments, many approaches leverage external resources (Al-Rfou et al., 2015). Cotterell and Duh (2017), for instance, harvest silver annotations from structured Wikipedia data and build models for typologically diverse languages, though their approach is limited to specific domains and label sets. Lin and Lu (2018) adapt well-resourced NER systems to low resource target domains, given minimal annotation and word embeddings in domain. Several translation-based approaches leverage better resourced languages by inducing lexical information from multi-lingual resources (Bharadwaj et al., 2016; Nguyen and Chiang, 2017; Xie et al., 2018). In a slightly different vein, Shang et al. (2018) use dictionaries as distant supervision to resolve entity ambiguity. Unfortunately, external resources are not always publicly available. It is in fact impossible to replicate many of the above studies without a government contract and extensive knowledge of linguistic resources, limiting their applicability to many DH scenarios. Mayhew et al. (2017) suggest manually building bilingual dictionaries when no other translation resources are available to facilitate their method, though active learning provides a more direct means of improving NER quality.

Active Learning Active learning seeks to maximize the performance of a model while minimizing the manual annotation required to train it. Shen et al. (2004) define three broad criteria for determining which data will be most *informative* to the model if annotated: *uncertainty*, where instances which confuse the model are given priority; *diversity*, where instances that would expand the model's coverage are prioritized; and *representativeness*, prioritizing instances that best approximate the true distribution over all instances. Uncertainty-based approaches outperform other single-criterion approaches, though many works, primarily in Computer Vision, demonstrate that considering diversity reduces repetitive training examples and representativeness reduces outlier sampling (Roy and McCallum, 2001; Zhu et al., 2003; Settles and Craven, 2008; Zhu et al., 2008; Olsson, 2009; Gu et al., 2014; He et al., 2014; Yang et al., 2015; Wang et al., 2018b).

For active learning in NER, Shen et al. (2017) propose the uncertainty-based metric maximized normalized log-probability (MNLP). It prioritizes sentences based on the length normalized log probability of the model's predicted label sequence. To make neural active learning tractable, they shift workload to lighter convolutional neural networks (CNN) and update weights after each manual annotation batch instead of retraining from scratch. They demonstrate state-of-the-art performance with MNLP, though Lowell et al. (2018) show its improvement above random sampling to be less dramatic, as do our experiments. Low-

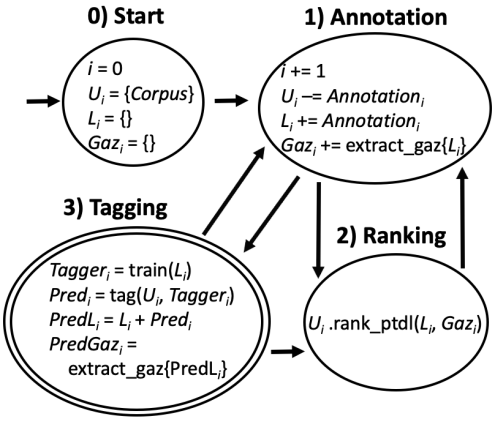

Figure 1: High level HER system architecture. Unlabeled sentences in U are manually labeled and moved to L, enabling iterative updates of gazetteers, the NER model, and the informativity ranking of sentences in U.

ell et al. (2018) compare calculating MNLP from the *native* inference model and from a *non-native* model with a separate architecture. They conclude that non-native models are ill-suited to active learning, which our findings using more robust informativeness criteria contradict.

3 The Humanities Entity Recognizer

As illustrated in Figure 1, HER consists of three components: (1) a human $User$ who provides an unlabeled corpus U at state 0 and annotates selected sentences in state 1, thus moving them from U to the labeled corpus L, (2) an active learning engine, $Ranker$, that ranks sentences from U in state 2 for $User$ to annotate based on how informative they might be to (3), the NER model, $Tagger$, to be trained on L in state 3.[2]

All states are linked by an interface that "white-boxes" the process. It advises $User$ on qualitative observations which might improve performance by manually interacting with $Ranker$, e.g., removing problematic gazetteer entries, or with $Tagger$, e.g., forcing it to sample some known minority labels. The contributions of the interface will not be reflected in our human-out-of-the-loop experiments on standard datasets, as these evaluate only the contributions of $Ranker$ and $Tagger$. Thus, reported performances should be considered a lower bound that can often be improved with minimal human intervention.

[2]In our experiments, we assume no previous annotation or pre-existing gazetteers at state 0, though, in practice, HER robustly leverages non-empty L_0 and/or Gaz_0 when available.

3.1 Ranking Sentences by Informativeness

At state 1 with i=0, $User$ is prompted to annotate randomly ordered sentences until 50-100 named entities are labeled. We use a 200-sentence seed for all experiments except that of Section 4.1, where an entity-sparse corpus requires a 300-sentence seed. Such a small seed, often annotated in less than 30 minutes, is sufficient to support $Ranker$'s *Pre-Tag DeLex* (PTDL) algorithm in state 2. PTDL uses only shallow, fast Conditional Random Fields (CRF) to avoid delaying manual annotation. As demonstrated on a sample corpus in Figure 2, PTDL involves four subtasks: *pre-tagging*, *delexicalized tagging*, *vocabulary weighting* and *sentence ranking*.

Pre-tagging We naively and greedily *pre-tag* U with binary entity–non-entity labels based on gazetteer matches. Hence, every n-gram in U matching a named entity from a gazetteer gets pre-tagged as such unless it overlaps with a longer named entity. State 2 cannot occur until the seed has been annotated, so Gaz will never be empty, as entities are automatically extracted into gazetteers after each annotation batch in state 1.

Delexicalized Tagging $U_{pre-tagged}$ is divided into U_{NE}, containing sentences with at least one pre-tagged named entity, and its complement, U_{noNE}. We train a *trusted* NER model (t) on L and two *biased* models (b_1 and b_2) on L plus random mutually exclusive halves of U_{NE}. b_1 and b_2 are biased in that they use non-gold data (U_{NE}) exhibiting an exaggerated density of named entities. Models are trained using only *delexicalized* features, which, unlike character n-grams for example, do not directly reference the focal word or its form. Many delexicalized features are context-based, like preceding and following words. Trained thus, models are less hampered by the class imbalance problem (Japkowicz, 2000), more likely to predict more named entities, and more capable of determining which Out Of Vocabulary (OOV) lexical items (in U but not L) make good named entity candidates.

Vocabulary Weighting After tagging U with delexicalized models, t, b_1, and b_2, OOVs are scored by weighted frequency. Weights are sums determined by which models tagged the OOV in an entity at least once. t contributes 1 to the sum and each biased model, $\frac{1}{2}$. OOVs not tagged by any model recieve a negligible positive weight, ϵ.

Pre-tagging

L	*Gaz*	
Sam/**PRS$_b$** likes Matt/**BND$_b$** and/**BND$_i$** Kim/**BND$_i$** and Bon/**BND$_b$** Iver/**BND$_i$**	**Persons**	**Bands**
	Sam	Bon Iver
		Matt and Kim

U	*U pre-tagged*
Sam loves Dina	Sam/**NE** loves Dina
Matt and Kim like Buicks	Matt/**NE** and/**NE** Kim/**NE** like Buicks
Dina loves Odesza	Dina loves Odesza

Delexicalized Tagging

U NE	*U noNE*	*U tagged*
Sam/**NE** loves Dina	Dina loves Odesza	Sam/**NE_t,b$_2$** loves Dina
Matt/**NE** and/**NE** Kim/**NE** like Buicks		Matt/**NE_t,b$_1$** and/**NE_t,b$_1$** Kim/**NE_t,b$_1$** like Buicks
		Dina/**NE_t,b$_1$,b$_2$** loves Odesza/**NE_b$_1$**

Vocabulary Weighting

OOV	Weighted Frequency
Dina	2*2 = 4
Odesza	0.5
loves	ε*2 = 2ε
Buicks	ε
like	ε

Sentence Ranking

U ranked	Score	OOVs Contributing to Score
Dina loves Odesza	1.5+0.7ε	Dina, loves, Odesza
Matt and Kim like Buicks	0.4ε	like, Buicks
Sam loves Dina	0	

Figure 2: Step-by-step example outputs from ranking unlabeled sentences in a sample corpus with PTDL.

This motivates PTDL to target frequent OOVs after exhausting OOVs more likely to be named entities.

Sentence Ranking As shown in Figure 2, sentences in U are ranked by the sum of scores of unique OOVs therein, normalized by the word length of the sentence. OOVs occurring in higher ranked sentences do not count toward this sum.

While typical active learning strategies for NER rely on the inference model's output probabilities, these are noisy, especially given scarce annotation. Data-scarce models lexically memorize training instances, yielding high precision at the expense of recall. They struggle to model non-lexical features more subtly correlated with entity status but also more likely to occur on OOVs. Hence, data-scarce models know what they know but are somewhat equally perplexed by everything else (Li et al., 2008). For this reason, uncertainty-based active learners can suffer from problematically weak discriminative power in addition to redundant and outlier-prone sampling.

By forcing reliance on delexicalized features and biasing models toward recall, our three-criteria approach identifies frequent (representativeness) OOV words (diversity) that are plausible candidate members of named entities. These make for better indicators of where the model may fail (uncertainty) because named entities are minority labels in NER and minority labels are challenging.

3.2 Sentence Tagging Architectures

$User$ can stop iteratively annotating and re-ranking U at any time to train a $Tagger$ on L to perform the full NER task on U (state 3). L is combined with $Tagger$'s predictions on U ($Pred$) to form $PredL$, from which an imperfect gazetteer is extracted ($PredGaz$). $User$ must inspect these to determine if additional annotation is required. We explore three $Tagger$ architectures:

CRF For tagging with Okazaki (2007)'s feature-based CRF, $Tagger$ first trains preliminary models on L, cross-validating on folds of the random seed. Each model leverages a unique permutation drawn from a universal set of features. The best performing feature set is used to train the final model. Training and inference are fast, even with preliminary cross-validation. In the exclusive evaluation, CRF is the best tagger until about 40K tokens of training data are acquired. In the inclusive evaluation, CRF's tendency to overfit is rewarded, as it outperforms both deep models regardless of corpus size.

CNN-BiLSTM The near state-of-the-art architecture proposed by Shen et al. (2017) aims to reduce training with minimal harm to accuracy. It leverages CNNs—as opposed to slower recurrent networks—for character and word encoding, and a bidirectional long short-term memory network (BiLSTM) for tags. CNN-BiLSTM outperforms

all other models in the exclusive evaluation for a stretch of the learning curve between about 40K tokens acquired and 125K. While faster than the other deep model considered here, training time is slower than the CRF and computationally costly.

BiLSTM-CRF The state-of-the-art BiLSTM-CRF architecture of (Lample et al., 2016) projects a sequence of word embeddings to a character level BiLSTM which in turn projects into a CRF at the tag level, with an additional hidden layer between the BiLSTM and CRF. In our experiments, BiLSTM-CRF surpasses CNN-BiLSTM performance once about 125K tokens are acquired.

3.3 HER in the Digital Humanities

HER was developed to benefit diverse DH projects. It is currently facilitating three such ventures.

The Herodotos Project (`u.osu.edu/herodotos`) aims at cataloguing ancient ethnogroups and their interactions (Boeten, 2015; de Naegel, 2015). HER is used to identify such groups in Classical Greek and Latin texts. Manually annotated data as well as a trained NER tagger are freely available from `github.com/alexerdmann/Herodotos-Project-Latin-NER-Tagger-Annotation`.

Artl@s `artlas.huma-num.fr` is a global database of art historical catalogs from the 19th and 20th centuries for the scholarly study of the diffusion and globalization of art. HER serves as a method for mining semi-structured texts characterized by noisy OCR and recurrent patterns of granular named entities.

Visualizing Medieval Places Wrisley (2017) concerns the study of recurrent places found within a mixed-genre corpus of digitized medieval French texts. NER has heretofore been challenged by sparsity from the unstandardized orthography. The related Open Medieval French project (`github.com/OpenMedFr`) benefits from HER's robust handling of sparsity, using the system to create open data regarding people and places referenced in medieval French texts.

4 Experiments

We now describe several experiments evaluating HER's performance on diverse corpora. When a standard test set is available, we perform inclusive evaluation on the combined train and dev sets and evaluate exclusively on test. Otherwise, we only evaluate inclusively. In both settings, we compare multiple combinations of ranking systems and taggers over a learning curve, reporting F1 exact match accuracy of identified entities. In all figures, line dashing (contiguous, dashed, or dotted) denotes inference model (CRF, BiLSTM-CRF, or CNN-BiLSTM), whereas line accents (stars, circles, triangles, or squares) denotes active learning method. Besides PTDL, we also consider a random active learning method (RAND), MNLP, and Erdmann et al. (2016)'s CAP algorithm. Like PTDL, CAP ranks sentences based on frequency weighted OOVs, but calculates weights based on capitalization patterns, prioritizing capitalized OOVs occurring in non-sentence initial position.

Quantity of training data is reported as percentage of the entire corpus for inclusive evaluations, and as tokens actively annotated (i.e., not counting the random seed sentences) for exclusive evaluations. For consistency, following seed annotation, we always fetch additional annotation batches at the following intervals, in tokens: 1K, 4K, 5K, 10K, 20K until we reach 100K total tokens, 50K until 300K total, 100K until 500K total, and 250K from there. For all experiments leveraging neural taggers, we use freely available pretrained embeddings (Grave et al., 2018), except for Latin, where we train fasttext (Bojanowski et al., 2017) embeddings on the Perseus (Smith et al., 2000) and Latin Library collections with default parameters (using pretrained embeddings yield small performance boosts that decrease with additional training data). We conclude this section with a direct comparison to the recently proposed active learning pipeline of Shen et al. (2017) and their MNLP ranking algorithm.

4.1 Consistency of Non-deterministic Results

Because the active learning pipeline involves taking a random seed and many of the experiments on larger corpora could not be averaged over several runs, we first measure performance variation as a function of ranking algorithm and quantity of annotation. Figure 3 displays our findings on a sample corpus of about 250K tokens[3] in five diverse, pre-1920 prose genres extracted from the FranText corpus (`www.frantext.fr`) and annotated for

[3] HER considers sentence boundaries to be tokens, as this helps users locate words, i.e., the line number will correspond to token number when rendered in CoNLL format.

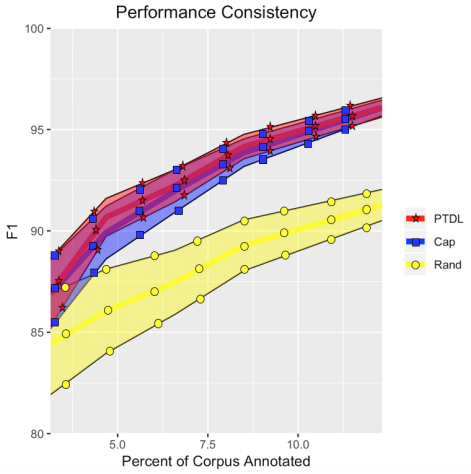

Figure 3: ± 1 standard deviation bands around the mean performance of each sentence ranking algorithm using the CRF tagger over 100 inclusive evaluations on our FranText corpus.

geospatial entities. Our sample covers topics from gastronomy to travel, exhibiting inconsistent entity density characteristic of DH corpora.

Noise is much higher for the first few batches of annotation, particularly due to the low recall of data scarce models. Reluctant to generalize, they behave more like look-up tables extracted from the seed, exacerbating the effect of random seed variation. After about 20K tokens annotated or 10% of the corpus, performance becomes much more predictable. All algorithms start with about a 5 point spread for ± 1 standard deviation, with means around 70 F, and all exhibit the diminishing variation trend, though RAND does less so. Unlike CAP and PTDL, subsequent annotation batches in RAND are not predictable from previous annotation batches. This results in a spread of 0.76 after annotating 12.33% of the corpus, whereas the other algorithms are close to 0.4.

While we only tested variation on one corpus, multiple runs on other corpora tended to reflect the same diminishing variation trends despite marked differences in entity granularity, density or corpus size. Switching to the exclusive evaluation only minimally increases variation. It was not feasible to rigorously test variation using neural taggers, though we note that they are somewhat more prone to seed related noise which does not diminish as rapidly as it does for CRF with more annotation.

In terms of performance, random annotation requires one to label between 23% and 31% of the corpus to achieve the performance of PTDL after labeling just 12%. For this corpus, PTDL reduces annotation time between 46% and 60%, requiring only 32K tokens from annotators instead of 60-80K. CAP's competitiveness with PTDL is not surprising given that French uses the capitalization standards it is designed to exploit. Both algorithms achieve 15% error reduction above RAND after the first post-seed annotation batch (left edge of Figure 3), increasing monotonically to 55% error reduction after the fifth batch (right edge).

4.2 Inclusive Versus Exclusive Evaluation

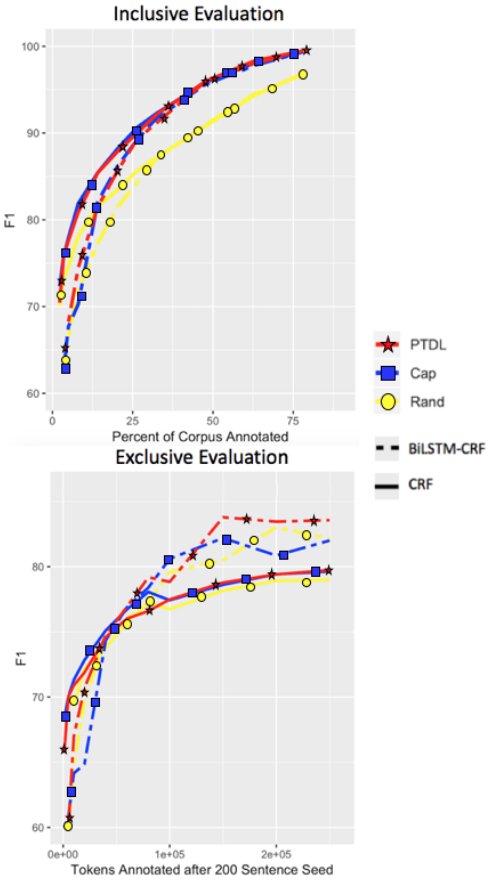

Figure 4: Comparing shallow and deep learning architectures on inclusive and exclusive evaluations with the CoNLL Spanish corpus.

Using the Spanish CoNLL corpus (Tjong Kim Sang and De Meulder, 2003) with canonical train, dev, test splits, we examine the relationship between evaluation framework, *Ranker*, and *Tagger* in Figure 4.[4] For the inclusive framework, *Ranker* selects sentences from train+dev for *Tagger* to train on, and the performance is calculated over the combination of those se-

[4] Lample et al. (2016) achieve 85.75 F on the exclusive evaluation, slightly beating our best BiLSTM-CRF models which sacrifice some performance for speed, switching to Adam optimization limited to 5 epochs.

lected sentences (trivially all correct) and trained $Tagger$'s predictions on the rest of train+dev. By reporting results over a learning curve, this evaluation framework is meaningful to the user whose primary goal is to produce a finite annotated corpus as efficiently and accurately as possible, a frequent concern in DH. The standard exclusive framework also gives $Ranker$ access to train+dev sentences, but calculates accuracy from $Tagger$'s predictions on the held out test set. The exclusive framework is thus more meaningful for future users of $Tagger$ who need the tool to generalize to sentences outside of train+dev.

In the inclusive framework, regardless of corpus size, BiLSTM-CRFs do not surpass CRFs until the accuracy is so high that the distinction is negligible. Promoting overfitting by reducing dropout did not significantly affect this result. In the exclusive framework, BiLSTM-CRF surpasses CRF around 50K tokens annotated. This holds for all languages and corpora we investigate, suggesting quantity of data annotated is more predictive of exclusive performance trends, whereas proportion of the corpus annotated better predicts inclusive trends.

4.3 Typology, Granularity, and Corpus Size

We consider the effect of language typology, label scheme granularity, and corpus size on inclusive and exclusive evaluations of taggers and rankers.

4.3.1 Insights from German

We repeat our experiments from Section 4.2 on the German NER corpus, GermEval (Benikova et al., 2014), to determine how robust our findings are to a larger corpus with finer label granularity and different capitalization standards. Our results in Figure 5 confirm many of our previous findings, with BiLSTM-CRFs overtaking CRFs of the same ranker after 50K tokens annotated on the exclusive evaluation. Shallow CRFs again dominate inclusively, and again, exclusive performance is less predictable, though the contribution of PTDL is more obvious.

GermEval contains over 520K tokens to Spanish CoNLL's 321K, showing that deep models are not just slower to overtake shallow models in the inclusive evaluation, but they only asymptotically approach shallow performance.[5] Furthermore, the finer grained named entity distinctions

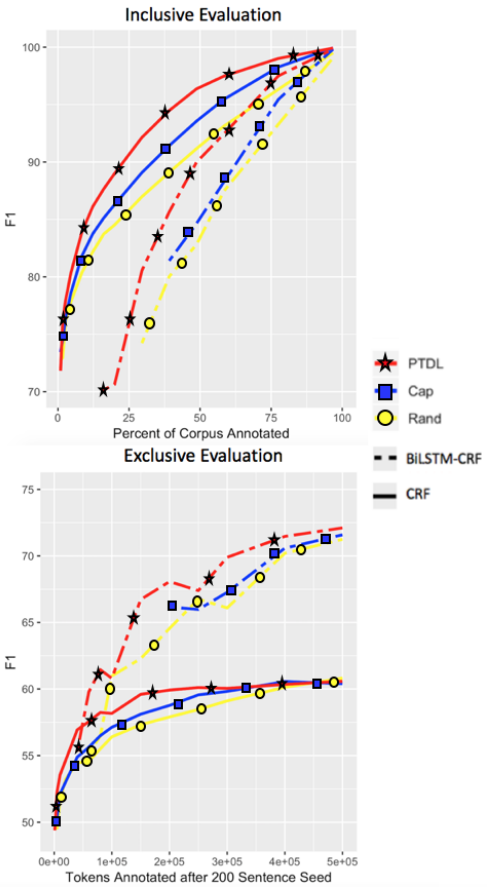

Figure 5: A comparison of shallow and deep learning architectures on inclusive and exclusive evaluations with the GermEval corpus.

in GermEval do not seem to affect our previous findings, but do cause BiLSTM-CRF to start slowly, as the model does not begin training until all possible labels manifest in the training set. While this is merely an effect of programming choices, it provides interesting insights. For instance, BiLSTM-CRF CAP models consistently start later than RAND which starts later than PTDL, meaning that PTDL is doing well on the diversity criteria, whereas CAP likely struggles because it relies on English-like capitalization standards. Since German capitalizes all nouns, CAP struggles here, having to search through many capitalized OOVs before finding named entities of each category. By not considering uncapitalized OOVs as named entity candidates, it can systematically avoid entire labels which do not take capitalization, such as dates. Thus, while PTDL performs robustly on the GermEval dataset, CAP is only weakly superior to RAND due to the weak correlation between entity status and capitalization.

[5] Our evaluation is equivalent to metric 3 from the shared task (Benikova et al., 2014), though our results are not comparable as we did not leverage nested labels.

4.3.2 Insights from Latin

Latin presents an opportunity to explore the impact of capitalization on ranking algorithms more thoroughly. Erdmann et al. (2016) selected their Latin corpus because English capitalization standards had been imposed during digitization, making CAP more likely to succeed. Figure 6 demonstrates that it even marginally outperforms PTDL on the corpus (left pane). However, capitalizing proper nouns is not a native attribute of Latin orthography and is not available in all digitized manuscripts, limiting the Latin texts in which CAP will succeed. The right pane in Figure 6 demonstrates this, as the same evaluation from the left pane is repeated on a lower cased version of the same corpus. The minuscule error reduction CAP achieves over RAND in this environment is due to its general preference for OOVs. Meanwhile, despite suffering from weaker named entity signals without capitalization, PTDL still manages to robustly identify what non-capitalization features are relevant, maintaining 25% error reduction over RAND. Finally, in German, where capitalization is a weak signal of entity status, PTDL is similarly better equipped to incorporate the weak signal, reducing error twice as much as CAP. Interestingly, PTDL performance in the lower cased Latin corpus almost exactly matches RAND performance on the capitalized version. This suggests the benefits of PTDL are comparable to the benefits of having English-like capitalization to mark entities.

4.3.3 Insights from Arabic

Unlike the other corpora, the news domain ANER Arabic corpus (Benajiba and Rosso, 2007) features rich templatic morphology, frequent lexical ambiguity, and an orthography lacking capitalization. Hence, not only will feature-based signals be more subtle, but the gazetteer-based pre-tagging component of PTDL will suffer from low precision, because Arabic is written in an abjad orthography where short vowels among other characters are seldom marked, making many words polysemous. Even so, PTDL significantly outperforms RAND, likely due to its ability to shift reliance to contextual features better suited for newswire, where formulaic expressions are often used to refer to certain entity types.

While PTDL compares well to RAND, it does not approach 100% accuracy after annotating 50% of the corpus as in Section 4.3.2. Besides ambiguity and lack of capitalization, this could be due to a typological bias in our "universal" feature set. Contiguous character n-grams, for example, will not capture non-concatenative subword phenomena. Going forward, we will investigate which feature sets were most useful as a function of language typology to identify gaps in our coverage.

4.4 Comparing to MNLP

Shen et al. (2017) and Lowell et al. (2018) evaluate the purely uncertainty-based MNLP active NER system on English corpora, reporting starkly different results. We address discrepancies and test the robustness of their findings by comparing MNLP to PTDL and RAND on the GermEval corpus. Results are displayed in Figure 8, with shaded regions corresponding to the range of performance over multiple runs. To compare fairly, we use the same CNN-BiLSTM tagger for all rankers and iteratively update weights instead of re-training from scratch after each annotation batch, as in Shen et al. (2017). We report results on our previously mentioned batch annotation schedule, though results were comparable using the batch schedule of Lowell et al. (2018). Shen et al. (2017) claim iterative updating does not affect accuracy significantly, though the best performing active CNN-BiLSTM in Figure 8 lags a few points behind the BiLSTM-CRF after 150K tokens annotated, with that gap reaching nearly 5 F when training on the whole corpus. Meanwhile, a CNN-BiLSTM trained from scratch on the whole corpus scores only 1 F less than the BiLSTM-CRF.

While Lowell et al. (2018) report improvement over RAND using MNLP when training on 0-10% of the corpus, we see little improvement after about 2%, and even then, PTDL greatly outperforms both. The relationship between the PTDL curves in the exclusive evaluation shows that CNN-BiLSTM is actually the optimal tagging architecture for a brief window, overtaking CRF around 30K tokens and staying in front of BiLSTM-CRF until about 125K tokens.

5 Conclusion and Future Work

We have presented the HER toolkit and its novel active learning algorithm, demonstrating robust handling of typological diversity, niche domains, and minority labels. The algorithm addresses the weak discriminative power of uncertainty-based models caused by class imbalance and precision bias. We also argued for the relevance of in-

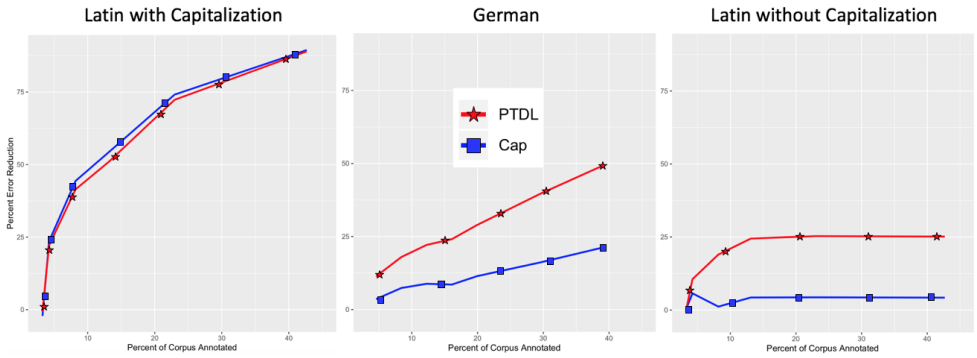

Figure 6: Percent error reduction over RAND in three corpora exhibiting typologically distinct capitalization standards. Corpora are presented in descending order of the correlation of capitalization with named entity status.

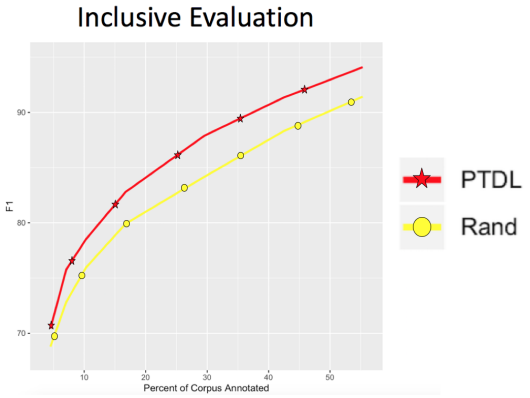

Figure 7: Effect of PTDL in the ANER Arabic corpus.

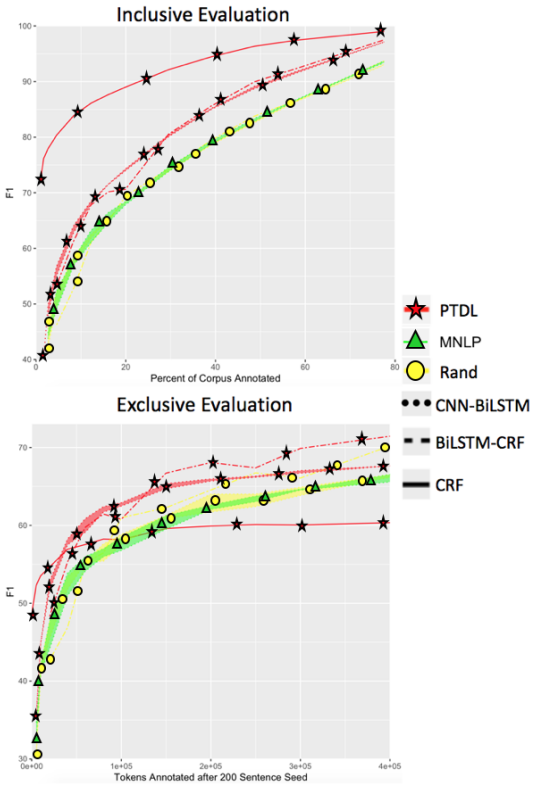

Figure 8: Inclusive and exclusive comparisons of the MNLP and PTDL rankers on GermEval.

clusive evaluations, demonstrating that a shallow CRF tagger outperforms deep taggers on sentences which the active learner could have selected for training. The CRF's tendency to overfit is rewarded in this case, as the selected sentences are especially representative of the remaining sentences to be tagged. When tagging held out test sets, CRFs are only optimal until about 30K training tokens are acquired, then CNN-BiLSTMs are preferable until 125K tokens when BiLSTM-CRFs become the best high resourced tagger.

In future work, we will investigate sources of noise in performance to see if these are due to gaps in the model, idiosyncrasies of corpora, or both. Additionally, we will expand HER to model hierarchically nested entity labels. Named entities are often difficult to label deterministically, inviting a problematic level of subjectivity, which is of crucial interest in DH and should not be oversimplified. We will consider strategies such as Wang et al. (2018a)'s shift-reduced-based LSTM architecture or Sohrab and Miwa (2018)'s method of modeling the contexts of overlapping potential named entity spans with bidirectional LSTM's.

Acknowledgments

We thank the Herodotos Project annotators for their contributions: Petra Ajaka, William Little, Andrew Kessler, Colleen Kron, and James Wolfe. Furthermore, we gratefully acknowledge support from the New York University–Paris Sciences Lettres Spatial Humanities Partnership, the Computational Approaches to Modeling Language lab at New York University Abu Dhabi, and a National Endowment for the Humanities grant, award HAA-256078-17. We also greatly appreciate the feedback of three anonymous reviewers.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-NER: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Yassine Benajiba and Paolo Rosso. 2007. Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information. In *Proceedings of Workshop on Natural Language-Independant Engineering*.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014. Germeval 2014 named entity recognition shared task: companion paper.

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.

Julie Boeten. 2015. The Herodotos project (OSU-UGent): Studies in ancient ethnography. Barbarians in Strabos geography (Abii–Ionians) with a case-study: the Cappadocians. Master's thesis, Gent Universiteit.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 91–96.

James R Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*, pages 164–167. Association for Computational Linguistics.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for Latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889.

Yingjie Gu, Zhong Jin, and Steve C Chiu. 2014. Active learning combining uncertainty and diversity for multi-class image classification. *IET Computer Vision*, 9(3):400–407.

Tianxu He, Shukui Zhang, Jie Xin, Pengpeng Zhao, Jian Wu, Xuefeng Xian, Chunhua Li, and Zhiming Cui. 2014. An active learning approach with uncertainty, representativeness, and diversity. *The Scientific World Journal*, 2014.

Nathalie Japkowicz. 2000. The class imbalance problem: Significance and strategies. In *Proceedings of the Intl Conf. on Artificial Intelligence*.

Kimmo Kettunen, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala, and Laura Löfberg. 2017. Old content and modern tools–Searching named entities in a Finnish OCRed historical newspaper collection 1771-1910. *DHQ: Digital Humanities Quarterly*, 11(3).

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Lihong Li, Michael L Littman, and Thomas J Walsh. 2008. Knows what it knows: a framework for self-aware learning. In *Proceedings of the 25th International Conference on Machine learning*, pages 568–575. ACM.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022.

David Lowell, Zachary C Lipton, and Byron C Wallace. 2018. How transferable are the datasets collected by active learners? *arXiv preprint arXiv:1807.04801*.

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.

Anke de Naegel. 2015. The Herodotos project (OSU-UGent): Studies in ancient ethnography. Barbarians in Strabos geography (Isseans–Zygi). with a case-study: the Britons. Master's thesis, Gent Universiteit.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 296–301.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing.

Nicholas Roy and Andrew McCallum. 2001. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the 18th International Conference on Machine Learning, 2001*, pages 441–448.

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.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 589. 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.

Rainer Simon, Elton Barker, Leif Isaksen, and Pau de Soto Cañamares. 2017. Linked data annotation without the pointy brackets: Introducing Recogito 2. *Journal of Map & Geography Libraries*, 13(1):111–132.

Rainer Simon, Leif Isaksen, ETE Barker, and Pau de Soto Cañamares. 2016. The Pleiades gazetteer and the Pelagios project. In *Placing Names: Enriching and Integrating Gazetteers*, pages 97–109. Indiana University Press.

David A Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *International Conference on Theory and Practice of Digital Libraries*, pages 127–136. Springer.

David A Smith, Jeffrey A Rydberg-Cox, and Gregory R Crane. 2000. The Perseus project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849.

Rachele Sprugnoli. 2018. Arretium or Arezzo? a neural approach to the identification of place names in historical texts. *http://www. ceur-ws. org*.

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-Volume 4*, pages 142–147. Association for Computational Linguistics.

Seth Van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. 2013. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*, 30(2):262–279.

Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin. 2018a. A neural transition-based model for nested mention recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1011–1017.

Zengmao Wang, Xi Fang, Xinyao Tang, and Chen Wu. 2018b. Multi-class active learning by integrating uncertainty and diversity. *IEEE Access*, 6:22794–22803.

David Joseph Wrisley. 2017. Locating medieval French, or why we collect and visualize the geographic information of texts. *Speculum*, 92(S1):S145–S169.

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.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1137–1144. Association for Computational Linguistics.

Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. 2003. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, pages 58–65. ICLM.