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 the Digital Humanities (DH) 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 humanists' 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 (NER), 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 NER for humanities corpora where such resources are not available. Evaluating on typologically disparate languages and datasets, we typically 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 in the Digital Humanities (DH). Humanists are increasingly interested in semantic annotation for under-resourced languages, imperfect or noisily structured data, and user-specific classification tasks, making it difficult to automate NER with black-box, off-the-shelf models. Manual annotation of large corpora from scratch, meanwhile, can be prohibitively costly. A successful digital initiative such as the Pelagios Commons (Simon et al., 2016), which collects geospatial data from historical sources, typically requires extensive funding, relying on 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, 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 jointly 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.

HER offers multiple architectures for the final NER model. We conduct extensive evaluation across many datasets in typologically diverse languages and domains to determine the best model as a function of the quantity of training data and intended test set. Despite the state-of-the-art performance of deep models on *exclusive*, held-out test sets (Lample et al., 2016), we demonstrate that shallow Conditional Random Fields (Lafferty et al., 2001) are preferable for *inclusive* test sets, i.e., sentences that were available to the sentence ranking algorithm during active learning. Whereas exclusive evaluation is assumed in traditional NER tasks, a finite test set is often intended from the outset in DH, making an inclusive evaluation more relevant. In an inclusive evaluation, we consider the entire corpus to be the test set, holding nothing out when performing active learning, as we are not concerned with generalizability, but how rapidly we can achieve the greatest cover-

[1] HER is freely available from github.com/alexerdmann/HER.

age of *that* corpus. Thus, the evaluation metric is calculated over both the manually annotated sentences—trivially all correct—and the remaining tagged sentences. Controlling for the inference 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), 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 data structure and noisy optical character recognition as input (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 jointly 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 represent variation in the data. 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 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. Lowell 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.

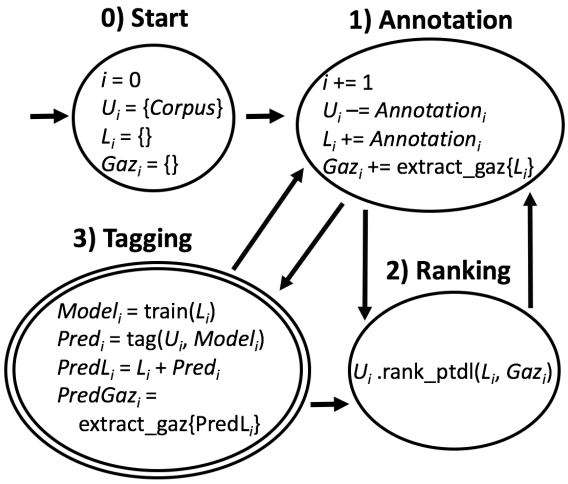

Figure 1: High level HER system architecture.

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 pre-existing labeled data and gazetteers if available), and batches of annotation in state 1, (2) an active learning engine, RANKER, that ranks sentences in U to be annotated based on how informative they might be to (3), the NER model, TAGGER, to be trained on L, the accumulation of USER's annotation in state 3. All states are linked by a humanist-friendly interface that "whiteboxes" 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 qualitative and quantitative benefits of the interface will not be reflected in our human-out-of-the-loop experiments leveraging previously annotated corpora, as these evaluate the independent, purely quantitative contributions of RANKER and TAGGER.

3.1 Ranking Sentences by Informativeness

At state 1, when $i = 0$, USER is prompted to annotate randomly provided sentences until 50-100 named entities are labeled. In all experiments, we use a 200-sentence seed, except for the entity-sparse French corpus, where we use 300. Such a small seed, often annotated in less than 30 minutes, is sufficient to support RANKER's *Pre-Tag DeLex* (PTDL) algorithm referenced in state 2.

PTDL uses only shallow CRFs, running in a few minutes to avoid delaying manual annotation. It begins by naively *pre-tagging* U based solely on longest match sequences with gazetteer entries. It then divides U into U_{NE}, containing only sentences with pre-tagged named entities, and its complement, U_{noNE}. Next, it trains a *trusted* model on L and two *biased* models on L plus random, mutually exclusive halves of U_{NE}. The models are trained to perform binary entity–non-entity labeling, using only *delexicalized* features, i.e., features that, unlike character n-grams for example, do not reference the focal word or its form. Trained thus, models are biased to expect higher entity densities, forced to consider non-lexical features at inference time, and less hampered by the class imbalance problem (Japkowicz, 2000).

Each OOV in U is scored according to weighted frequency, where weights are sums determined by which models tagged the OOV in an entity at least once. The trusted model contributes 1 to the sum and each biased model, .5. A negligible positive weight is applied when no model tags the OOV in an entity, as this motivates the algorithm to focus on frequent OOV words once it exhausts candidate entity OOVs. Finally, sentences in U are ranked in descending order by the length normalized sum of scores of unique OOVs therein which do not occur in any higher ranked sentence.

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 named entity candidates. These are better indicators of where the model may fail because named entities are minority labels in NER and minority labels are challenging (uncertainty).

3.2 Sentence Tagging Architectures

USER can stop iteratively annotating and re-ranking sentences at any time to train a model on L to perform the full NER task on U (state 3). L and the tagged U are combined into a fully labeled corpus, $PredL$, and a gazetteer is extracted, $PredGaz$. USER qualitatively inspects these to determine if additional annotation is required. Three tagging models are available:

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 40K tokens acquired and 100-150K. 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 by the time 150K 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. 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., 2016) embeddings on the Perseus (Smith et al., 2000) and Latin Library collections with default parameters. In experiments not reported here, we get small boosts

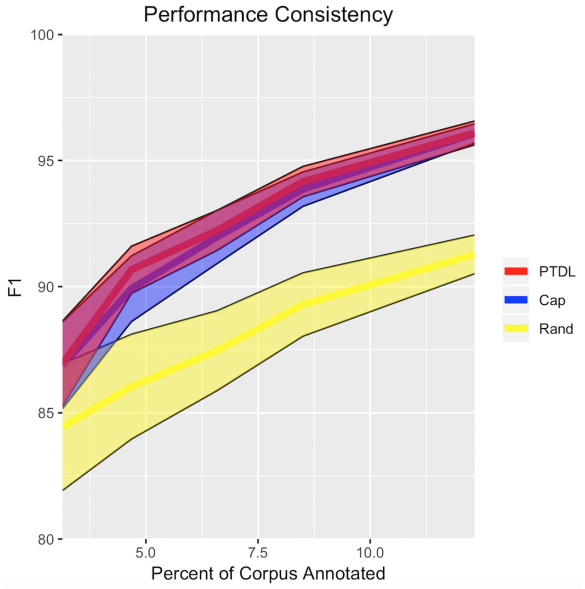

Figure 2: ± 1 standard deviation bands around the mean performance of each sentence ranking algorithm over 100 inclusive evaluations using the CRF tagger.

from pretrained embeddings that decrease with additional training data. Results are reported in the CoNLL format, i.e., F1 exact match accuracy of identified entities. In addition to PTDL, we also consider RAND, a random sentence ranker, and CAP, the capitalization dependent algorithm described in Erdmann et al. (2016). Like PTDL, CAP also ranks sentences based on frequency weighted OOVs, but calculates weights based on capitalization patterns, prioritizing capitalized OOVs occurring in non-sentence initial position. 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 2 displays our findings on a sample corpus of about 250K tokens[2] in five diverse, pre-1920 prose genres extracted from the FranText corpus (`www.frantext.fr`) and annotated for geospatial entities. Our sample covers topics from gastronomy to travel, exhibiting inconsistent entity density characteristic of DH corpora.

[2] 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.

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 .76 after annotating 12.33% percent of the corpus, whereas the other algorithms are close to .4.

While we only tested variation extensively in this corpus using the CRF tagger, any multiple runs we conducted on other corpora were vaguely consistent with the levels of variation reported here, despite marked differences in entity granularity, density and corpus size. Switching to exclusive evaluation only minimally increases variation. It was not feasible to run a comparably extensive study of variation using neural taggers, though we note that they are somewhat more prone to seed related noise which does not diminish with more annotation as rapidly as it does for CRF.

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 2), increasing monotonically to 55% error reduction after the fifth batch (right edge).

4.2 Internal Versus External Evaluation

Using the Spanish CoNLL corpus (Tjong Kim Sang and De Meulder, 2003) with canonical train, dev, test splits, we examine the relationship between evaluation setting and tagger choice. Figure 3 displays our results. Lample et al. (2016) report 85.75 F on the exclusive evaluation, slightly beating our best BiLSTM-CRF models which sacrifice some accuracy for speed, switching to

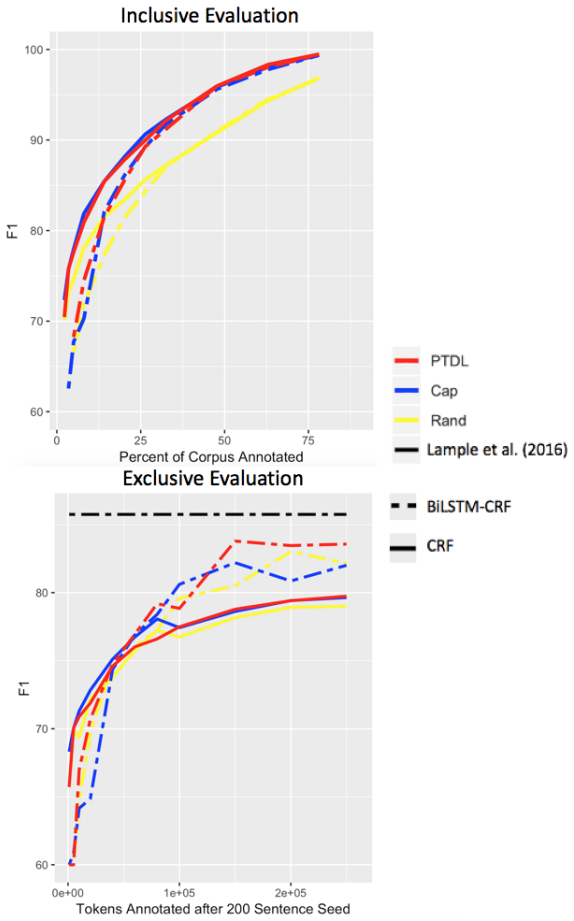

Figure 3: A comparison of shallow and deep learning architectures on inclusive and exclusive evaluations with the CoNLL Spanish corpus.

Adam optimization limited to 5 epochs.

Interestingly, no matter how many tokens are provided by any ranking algorithm, BiLSTM-CRFs only approach and never surpass the performance of CRFs in the inclusive evaluation. Additional experiments reducing dropout did not affect this result. While performance is less predictable in the exclusive evaluation, BiLSTM-CRF always surpasses CRF near 50K tokens annotated. This holds relatively true 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 trends in inclusive evaluation.

4.3 Typology, Granularity, and Corpus Size

In this section, we consider the impact of language typology, granularity of label schemes, and corpus size, on both the inclusive and exclusive evaluations of taggers and sentence rankers.

4.3.1 Insights from German

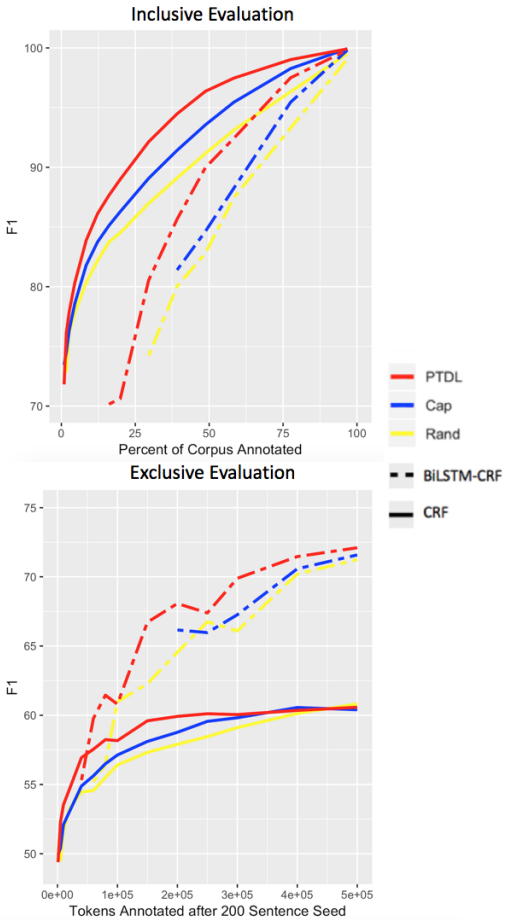

Figure 4: A comparison of shallow and deep learning architectures on inclusive and exclusive evaluations with the GermEval corpus.

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 4 confirm many of our previous findings, with BiLSTM-CRFs overtaking CRFs of the same ranker after 50K tokens annotated on the exclusive evaluation. Shallow models 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.[3] Further-

[3]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.

more, the finer grained named entity distinctions 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.

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 5 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 5 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 accuracy in the lower cased Latin corpus is on average only .16 F below RAND on the capitalized version of the same corpus. 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, ubiquitous lexical ambiguity, and is written in an orthography which does not express 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 removed, making many words polysemous. Even so, PTDL still 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 the Latin corpus. While this could be due to high ambiguity and lack of capitalization, it could also signal typological bias in our feature set. Contiguous character n-grams, for example, will not capture non-concatenative subword relationships. In ongoing work, we are investigating which feature sets were most useful as a function of language typology and quantity of training data 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 7, 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 we produced comparable results 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 7 lags a few percent behind the BiLSTM-CRF after 150K tokens

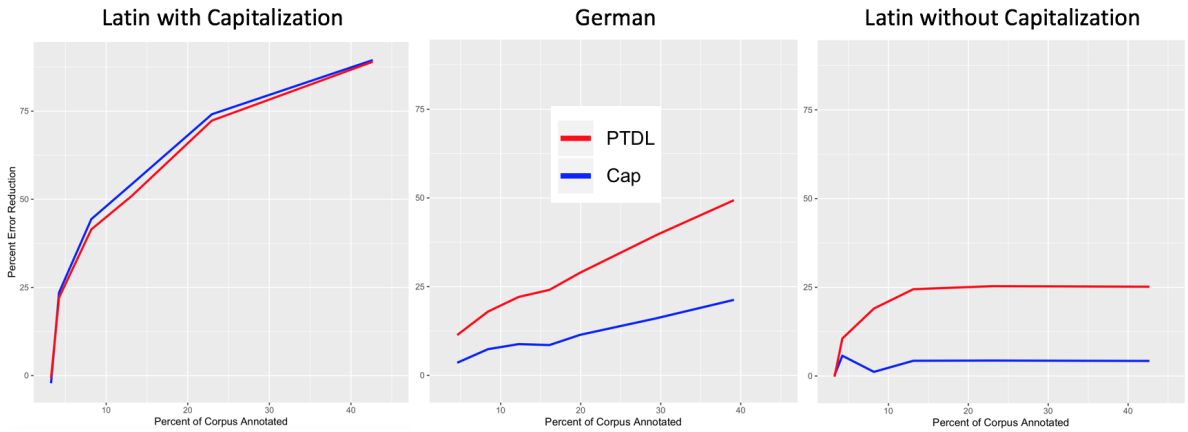

Figure 5: Percent error reduction over RAND in three corpora exhibiting typologically distinct capitalization standards. Corpora are presented in descending order of the correlation of capitalization with named entity status.

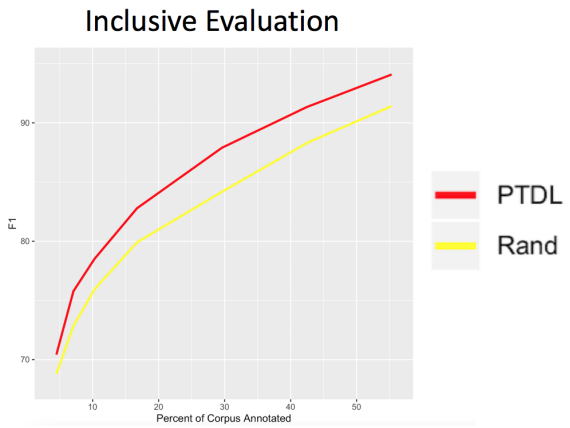

Figure 6: Effect of PTDL in the ANER Arabic corpus.

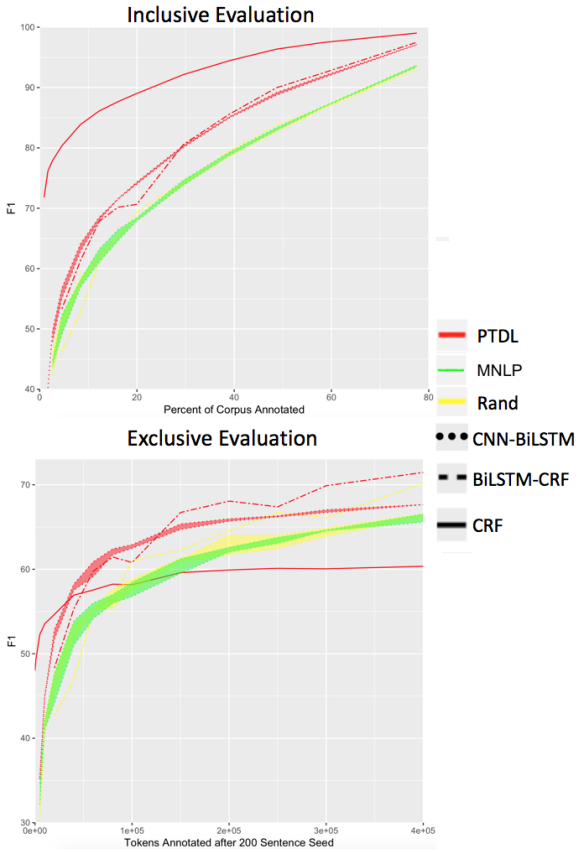

Figure 7: Inclusive and exclusive evaluations of ranking algorithms on GermEval, pitting the recently proposed MNLP algorithm against PTDL.

annotated, with that gap reaching nearly five percent when training on the whole corpus. Meanwhile, a CNN-BiLSTM trained from scratch on the entire corpus scores only one percent 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 is useful for determining optimal tagger architecture, as it shows that the CNN-BiLSTM is actually optimal for a brief window, overtaking the shallow CRF around 30K tokens and staying in front of the BiLSTM-CRF until over 100K tokens.

5 Conclusion and Future Work

We have described a novel active learning system, HER, and demonstrated its utility in DH, as it robustly handles typologically diverse languages, low resource environments, and minority labels. We made theoretical contributions as well, arguing for the importance of inclusive evaluations and identifying the weakened discriminative power of low resource uncertainty based models due to class imbalance and precision bias.

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 gratefully acknowledge insightful conversations with the Herodotos Project Latin and Greek annotation team: Petra Ajaka, William Little, Andrew Kessler, Colleen Kron, and James Wolfe. Furthermore, we are indebted to the support of 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. Finally, we wish to thank three anonymous reviewers for their feedback.

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. Anersys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information. In *IICAI*, pages 1814–1823.

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. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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 *Proc. 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. *arXiv preprint arXiv:1603.01360*.

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. *IJCNLP 2017*, page 296.

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 Monte Carlo estimation of error reduction. *ICML, Williamstown*, 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, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. *arXiv preprint arXiv:1809.03599*.

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 *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, volume 3.