ELLEN: E*xtremely* L*ightly Supervised* L*earning For* E*fficient* N*amed Entity Recognition*

Haris Riaz, Razvan-Gabriel Dumitru, Mihai Surdeanu
University of Arizona
Tucson, AZ, USA
{hriaz, rdumitru, msurdeanu}@arizona.edu

Abstract

In this work, we revisit the problem of semi-supervised named entity recognition (NER) focusing on extremely light supervision, consisting of a lexicon containing only 10 examples per class. We introduce ELLEN, a simple, fully modular, neuro-symbolic method that blends fine-tuned language models with linguistic rules. These rules include insights such as "One Sense Per Discourse", using a Masked Language Model as an unsupervised NER, leveraging part-of-speech tags to identify and eliminate unlabeled entities as false negatives, and other intuitions about classifier confidence scores in local and global context. ELLEN achieves very strong performance on the CoNLL-2003 dataset when using the minimal supervision from the lexicon above. It also outperforms most existing (and considerably more complex) semi-supervised NER methods under the same supervision settings commonly used in the literature (i.e., 5% of the training data). Further, we evaluate our CoNLL-2003 model in a zero-shot scenario on WNUT-17 where we find that it outperforms GPT-3.5 and achieves comparable performance to GPT-4. In a zero-shot setting, ELLEN also achieves over 75% of the performance of a strong, fully supervised model trained on gold data. Our code is publicly available.

Keywords: semi-supervised learning, named entity recognition, neuro-symbolic, rules, language models, modular architectures

1. Introduction

Named entity recognition (NER), i.e., the task of identifying named (and sometimes numeric) entities such person and organization names, drugs, protein names, diseases, and dates, is one of the earliest formal natural language processing (NLP) tasks (Grishman and Sundheim, 1996). NER remains critical to many real-world applications such as question answering and information extraction (Yadav and Bethard, 2019). Despite the tremendous progress observed on the NER task in the past almost three decades, we argue that there are several practical limitations in the way this task is generally formalized, which impact our understanding of what methods perform best in practice. In particular:

(1) Current settings for the NER task require an amount of annotations that are unrealistic for many real-world applications. For example, a common setting for semi-supervised NER uses 5% of the CoNLL-2003 corpus' (Tjong Kim Sang and De Meulder (2003)) training data, or over 10K total tokens (Chen et al., 2020; Zheng et al., 2023). In our work (Vacareanu et al., 2024), we have observed that NER annotations take approximately 3.2 seconds per token in practice. Thus, annotating the equivalent amount of data in a new domain would take approximately 9 person hours. This is unrealistic in many scenarios (e.g., intelligence, pandemic surveillance) that require the rapid development of custom models and where domain experts "do not want to come willingly and do not come cheaply."[1]

(2) While recent directions that use in-context learning (ICL) for NER with autoregressive decoder-based large language models (LLMs) perform well (Chen et al., 2023), they do not scale as well as encoder-based methods due to the decoder's high inference overhead; each generated token requires its own forward pass through the model.

(3) Recent trends rely mostly on neural networks (NNs) to learn the NER task, ignoring linguistic hints such as "one sense per discourse" (Gale et al., 1992) that might be present and are likely to be useful in lightly-supervised settings.

To remedy these limitations we propose an *extremely lightly supervised* scenario for NER, in which the only supervision comes in the form of a lexicon containing 10 examples per entity class. Importantly, the 10 examples are selected by a domain expert that does *not* have access to any dataset annotations. Further, we propose a simple NER approach for this scenario that is efficient and performs well despite the limited supervision. Our method uses an encoder-only inference strategy, but, at training time, it combines multiple strategies including language models and several linguistic heuristics. We call our method E*xtremely*

[1] IARPA program manager, personal communication

LREC-COLING 2024, pages 5622–5636

L*ightly Supervised* L*earning for* E*fficient* N*amed Entity Recognition* (ELLEN)[2].

Our main contributions are as follows:

(1) We demonstrate the effectiveness of combining language models with commonsense linguistic rules inspired by (Liao and Veeramachaneni, 2009) and aggregated under a self-training, modular, neuro-symbolic architecture. Our approach is considerably simpler than other complex statistical methods for semi-supervised NER (Nagesh and Surdeanu, 2018; Lakshmi Narayan et al., 2019; Peng et al., 2019; Zhou et al., 2022; Chen et al., 2019; Clark et al., 2018; Chen et al., 2020; Zheng et al., 2023, inter alia).

(2) Our approach includes a novel component called the Masked Language Modeling (MLM) Heuristic, which is a fully unsupervised NER method that achieves over 55% precision on the CoNLL-2003 NER dataset. Further, this component complements other self training as well as linguistic heuristics in the semi-supervised setting.

(3) We evaluate our method on CoNLL-2003 (Tjong Kim Sang and De Meulder (2003)) under three different degrees of supervision, and in a zero-shot setting on WNUT-17 (Derczynski et al. (2017)). On CoNLL, under the proposed setting of extremely limited supervision, we show that our method achieves an F1 score of **76.87**%. Further, when we increase the degree of supervision to match other methods which are state-of-the-art in the semi-supervised NER setting, we find that our method achieves comparable performance. We also show that our method continues to scale, even when using all of the data available for supervision. In a zero-shot evaluation on WNUT-17, we find our method to be comparable to LLMs such as GPT-3.5 (OpenAI, 2023b) and GPT-4 (OpenAI, 2023a), while also obtaining over 75% of the performance of a fully supervised model trained on WNUT-17.

2. Related Works

Recently, Large Language Models (LLM's) have emerged as the dominant approach for a wide variety of NLP tasks, including Named Entity Recognition. (Wang et al., 2023) and (Zhou et al., 2023) show that LLM's consistenly achieve SOTA performance on many NER datasets. With In-Context Learning, LLM's have also proven to be very useful in the FewShot NER setting, as recently shown by (Ashok and Lipton, 2023). However, LLM's typically have a high inference overhead (Narayanan et al., 2023) and may not always perform well in specialized or low-resource domains. Moreover, there are increasing concerns about data contamination. (Golchin and Surdeanu, 2023) demonstrate that GPT-3.5 and GPT-4 have encountered test data with labels from widely-used NLP benchmark datasets during pre-training. (Sainz et al., 2023) claim that this is true for CoNLL-2003, one of our evaluation datasets.

Focusing on Semi-Supervised NER, and not FewShot NER, current state-of-the-art methods include JointProp (Zheng et al., 2023), which is a multi-task learning framework that jointly tries to solve relation classification and NER using a heterogeneous graph structure. Semi-LADA (Chen et al., 2020) adapts the mixup data augmentation technique to sequence labeling, and then trains on linearly interpolated pairs of samples. Both of these methods use at least 5% of the labeled data as their most minimally supervised setting, which we argue, is an impractical level of supervision for semi-supervised NER. Another class of statistical methods (Peng et al., 2019; Zhou et al., 2022) try to solve the reliance on gold labels by resorting to distant supervision: they construct lexicons based on large dataset independent knowledge bases. These methods then use Positive-Unlabeled (PU) learning to train classifiers using only labeled positive examples and a set of unlabeled data containing both positives and negatives.

Other approaches like (Liu et al., 2019a) showed the benefits of augmenting neural NER taggers with external gazetteers. However, external gazetteers may not always be available for particular domains. In contrast we propose a semi-supervised method for NER along the lines of the approach taken by (Liao and Veeramachaneni, 2009): one that combines the generalizability of contemporary deep learning with intuitive reasoning and linguistic insights, and we demonstrate its strong performance, under a setting of extremely low supervision. We posit that such an approach can rival more multi-faceted statistical techniques for semi-supervised NER, such as PU learning & data augmentation, among others.

3. Proposed Method

As our method does not rely on any explicitly labeled texts, we begin by discarding all labels in the NER dataset (CoNLL-2003 in this paper). We then ask a domain expert to generate a small lexicon of 10 example named entities per class, for each of the four classes in CoNLL-2003, i.e., `PER`, `ORG`, `LOC`, and `MISC`. This lexicon: a) is sourced entirely from the tokens in the dataset; b) is constructed *without* looking at any of the labels in the dataset; c) does not rely on any external knowledge-base or dictionary; and d) serves as the sole source of "gold supervision" for our method. The domain expert is able to construct the lexicon (refer to Table 1) in less than 30 minutes for CoNLL-2003. We believe

[2] https://github.com/hriaz17/ELLEN

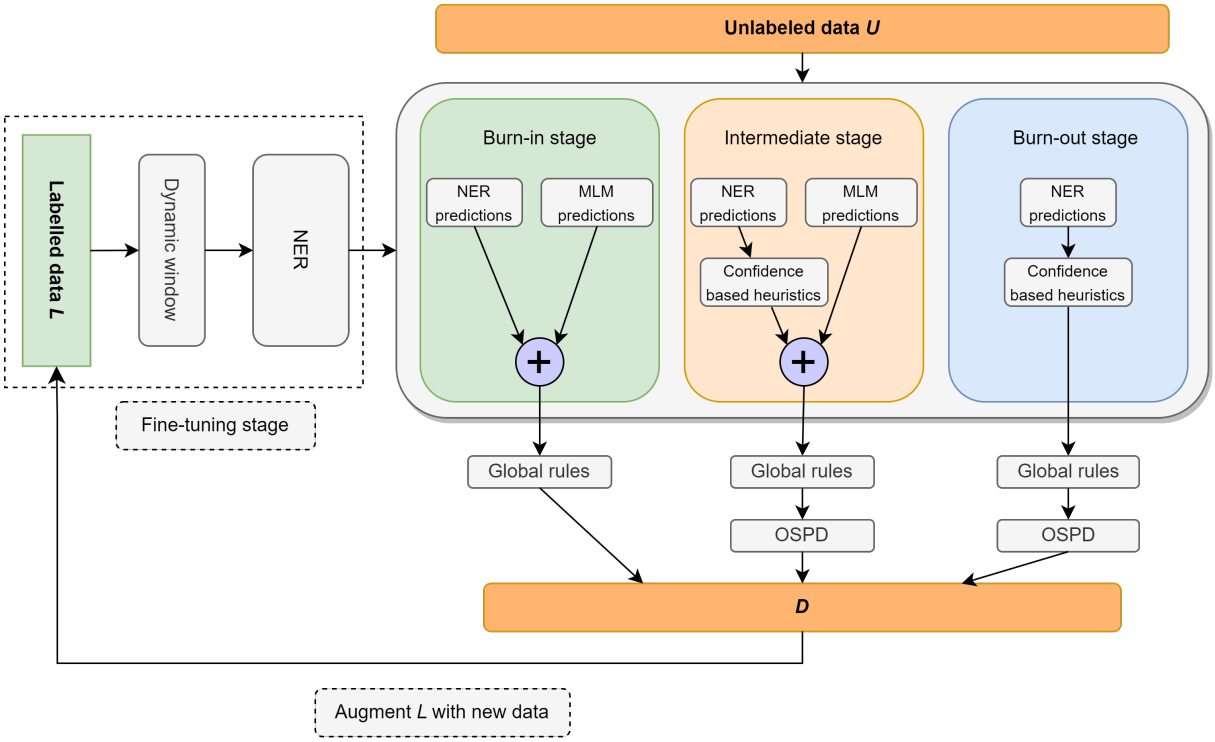

Figure 1: The proposed method illustrated. D refers to a subset of the unlabeled data which is added back to the labeled data for retraining in the next iteration. OSPD refers to the "One Sense Per Discourse" rule; "Global rules" indicate the rules described in Section 3.3. The colors used in the figure represent the decreasing quality of the generated annotations in the three stages, after the fine-tuning stage: green → orange → blue.

Category	Entities
ORG	Reuters, PUK, NATO, Honda, Ajax Amsterdam, Motorola, PSV Eindhoven, PKK, Hansa Rostock, Commonwealth
LOC	Germany, Australia, Britain, Spain, Italy, LONDON, Russia, China, Japan, NEW YORK
MISC	Dutch, British, French, Russian, German, Iraqi, Israeli, English, Australian, American
PER	Clinton, Yeltsin, Arafat, Lebed, Wasim Akram, Waqar Younis, Mushtaq Ahmed, Netanyahu, Williams, Rubin

Table 1: Lexicon generated by the domain expert following the criteria that we outline in Section 3.6. This lexicon serves as the seed set of entities for our model and the sole source of "gold" supervision.

10 examples per class is a reasonable number for producing an informative lexicon with the minimum amount of effort required; this is similar to few-shot relation extraction, which focuses on 5 examples/5-shots (Han et al., 2018). We then use this lexicon to annotate a small portion of the entirely unlabeled data of CoNLL-2003. In terms of "degree of supervision" (refer to table 4), we annotate about 9.13% of entities present in the CoNLL training data using the lexicon. We refer to this annotated subset as L and the remaining unlabeled data as U, following the convention of traditional semi-supervised learning literature. At a high level, our approach extends the simple self-training algorithm presented by (Liao and Veeramachaneni, 2009), which is shown in Algorithm 1. However, we argue that the procedure delineated in this algorithm is abstract, offering no guidance on the precise sequence in which the NER classifier C_k (and any other linguistic rules) should be applied to extract new data D. Such ambiguity could potentially lead to the common pitfalls inherent in classic self-training approaches. In our study, determining the optimal order and manner in which these various approaches should be combined proved to be a nuanced task. Our work is motivated by curriculum learning, which argues that better models are learned when the training data is "presented in a meaningful order which illustrates gradually more complex examples" (Bengio et al., 2009; Sachan and Xing, 2016). In this paper, we adjust this principle to mean *gradually noisier examples* based on the observation that in self-training it is critical that the initial models

Algorithm 1 A simple NER self-training algorithm

Given:
  L - a small set of labeled training data
  U - unlabeled data
for k iterations **do**
  Step 1: Train a NER C_k based on L
  Step 2: Extract new data D based on C_k
  Step 3: Add D to L
end for

be of higher quality to reduce noise in future iterations. Inspired by this idea, we propose an intuitive, three-stage framework, illustrated in Figure 1, for effectively combining linguistic rules with pre-trained language models such as (He et al., 2023). We design our framework to simultaneously balance two orthogonal goals:

1. avoiding the pitfalls of classic self-training (e.g., a model failing to correct its errors and instead amplifying them) to the extent possible, and

2. still being conceptually similar to self-training at a high level.

Our proposed NER method is fully modular, uses the `deberta-v3-large` encoder[3] as the neural component, and blends various other linguistic and statistical heuristics in a sieve (Lee et al., 2013). We first describe each of these heuristics below, and then describe how they are integrated in the three-stage ELLEN framework.

3.1. Unsupervised Entity Recognition Using A Masked Language Model (MLM)

Motivated by the observation that any language model, subjected to pretraining via the Masked Language Modeling (MLM) objective, likely acquires semantic, syntactic, and world knowledge, we hypothesize that the capability to discern named entities is also inherently embedded within such models. We present a novel, fully unsupervised algorithm, implemented as a rule in our neuro-symbolic modular architecture that allows us to gain additional "free" supervision , beyond our small lexicon of ten entities. This algorithm relies on a small pre-trained LM from (Liu et al., 2019b), which leverages our lexicon to extract new Named Entities from unlabeled data.

Unlike recent prompt-based approaches for NER, particularly few-shot NER, which involve either prompting LLM's with in-context examples to inject NER ability (Chen et al., 2023), or involve constructing dynamic templates based on label aware pivot words, our approach is much simpler and more constrained. We first use a very simple, linguistically inspired regular expression, based on part-of-speech (POS) tags, for detecting named entity spans:

```
(NNP|NNPS)+(IN(NNP|NNPS)+)?
```

where `NNP`/`NNPS` are the POS tags of singular/plural proper nouns, and `IN` is the POS tag assigned to prepositions. On the unlabeled portion of the CoNLL training dataset, this rule can detect named entity boundaries with a precision of **85.16%**, as shown in Table 2. Note that this rule is considerably simpler and more efficient than other recent computationally-intensive approaches, e.g., the entity typing and span identification method of Shen et al. (2023), or the span classification approach adopted by Arora and Park (2023) and Chen et al. (2022). Typically, these approaches initially train a model for span classification, followed by a model for entity type classification.

Precision	Recall	F1 Score
85.16%	90.96%	87.96%

Table 2: Micro-F1 scores of the regex rule for detecting named entity boundaries.

Once entity spans are identified, we label them using a masking heuristic. Intuitively, our method selects the entity label whose exemplars in the lexicon fill in the span with the highest likelihood. More formally, we first mask the span identified by the regex above with a number of `[MASK]` tokens equal to the tokens included in the span. For example, the span *John Doe* in the sentence: *John Doe is happy* will be masked as `[MASK] [MASK]` *is happy*. We then iteratively fill in lexicon entries of the same length (across all entity classes) and keep track of all token probabilities. For example, the entry *Dole* in the Person lexicon, which is tokenized as *Do* and *##le*,[4] produces the sentence: *Do ##le is happy*. Lastly, we select the entity label based on the following formula:

$$c = \arg\max_k \max_j \frac{1}{n} \sum_i^n p(t_i|x)$$

where t_i is the ith masked token (e.g., *Do* is t_0 in the example above); x is the sentence with the masked tokens; n is the total number of masks for the current example (e.g., 2 for the example above); the

[3] DeBERTa-v3 is the current state-of-the-art encoder-based model on many benchmark NLP tasks. The key advantage of using DeBERTa is its relative positional encoding, which allows the model to generalize better to longer sequences.

[4] We use the BERT tokenizer convention for multi-token words here for readability.

j index iterates over all exemplars for the current entity label; and k iterates over all entity class lexicons. That is, for each exemplar, we first compute the average probability of all its tokens. Then we pick the exemplar with the highest probability in a given lexicon as the probability of the corresponding entity label. Lastly, we select the entity label c with the highest probability.

We denote this rule as the Masked Language Modeling (MLM) Heuristic. We demonstrate its NER effectiveness on the development set of CoNLL-2003, in a fully unsupervised fashion in Table 3. As shown in the table, the F1 score of the MLM heuristic is over 56% on the development set of CoNLL-2003. In each iteration of the procedure shown in figure 1, the MLM is used in the burn-in and intermediate stages to annotate a subset of the unlabeled data U, which is eventually added back to L.

Entity Type	Precision	Recall	F1
Overall	61.78%	51.90%	56.41%
LOC	69.72%	41.53%	52.05%
MISC	45.18%	55.15%	49.67%
ORG	44.85%	40.88%	42.77%
PER	85.07%	65.02%	73.71%

Table 3: P/R/F1 of the Masked Language Modeling Heuristic as a fully unsupervised NER algorithm on the CoNLL-2003 development set. It obtains an entity-level F1 score of 56.41% with over 60% overall precision.

3.2. Dynamic Window Filtering

In most lightly supervised settings, NER models tend to suffer from the "unlabeled entity problem" as described in (Li et al., 2021), where the entities of a sentence may not be fully annotated. This tends to seriously degrade model performance, since the model treats unlabeled entities as negative or O/Outside instances. Even self-training methods may not be sufficient to completely alleviate the false negative problem since they are susceptible to confirmation bias (Arazo et al., 2020), i.e., erroneously predicted pseudo-labels are likely to deteriorate the model's performance in subsequent rounds of training. In contrast to (Li et al., 2021)'s method which uses negative sampling to avoid training the NER on unlabeled entities, we propose a very simple and run-time efficient linguistically inspired algorithm for controlling the effect of false negatives in sparsely annotated data settings, such as ours. We refer to our algorithm as "Dynamic Window Filtering." Using part-of-speech (POS) tag information[5] and an intuition that tokens which are labeled as O/Outside and which possess a POS tag of NNP (singular proper noun) are highly likely to be unlabeled named entities and thus should be discarded from the NER's training data. We implement this algorithm as the following rule: we slide a contextual window across each sentence in the labeled subset of the data, L, and for each named entity segment we encounter whose label is known, we create a window of size W, which dynamically expands in both directions around the labeled entity until an O token that is also tagged with the POS tag of NNP (singular proper noun), is encountered[6]. We also emphasize that our POS tags are inherently noisy, since they are obtained from an external LSTM-CRF based POS tagger that was trained exclusively on the Penn Treebank corpus (Marcus et al., 1993). We *do not* use the gold POS tags from the CoNLL data. The example below illustrates this rule:

Example 1:
EU *rejects* ***German*** *call to boycott* ***British*** *Lamb.*

Suppose "EU" and "British" are both Named Entities with known labels and are also proper nouns. Suppose "German" is also a proper noun but with an unknown named entity label (and, thus, it is currently labeled as O). Dynamic Window Filtering creates a contextual window around "EU" and "British", expanding in size until it encounters the token "German". This algorithm would thus break the original example into two new segments:

1. "**EU** rejects"
2. "call to boycott **British** lamb."

In every stage of our method, we apply dynamic window filtering on the data that the NER is trained on. This includes both the initial set of sparsely annotated gold data and its augmentations with the pseudo-labeled data (D) that is extracted from the unlabeled data (U) in each iteration. We find that this algorithm achieves the same goal as the method presented in (Li et al., 2021) i.e. discarding Named Entities with unknown labels, while being much simpler and computationally cheaper.

[5]POS tags are now available for many languages (see https://universaldependencies.org/) and can be obtained from various off-the-shelf models or language processing tools.

[6]We note that alternatives to Dynamic Windowing exist for managing unlabeled entities. For instance, as demonstrated in (Vacareanu et al., 2024), unlabeled entities can remain in the NER's training data with their impact mitigated by excluding them from the loss calculation, i.e., by backpropagating only over gold-annotated tokens.

3.3. Global Rules

Lightly supervised NER models may confuse named entities between Organizations and Persons, Organizations and Locations (and vice versa) due to their shared context. To remedy this, we apply a series of commonsense linguistic rules on the aggregated predictions of the NER model and the Masked Language Modeling (MLM) heuristic. We apply rules for disambiguating named entity segments that have been tagged as Persons (PER) but end in a company suffix. We update the labels of such segments, including the company suffix to ORG. For example, if the entity segment "Walt Disney" is tagged as PER, but it is immediately followed by "Inc." (a company suffix), the rule would force the whole segment "Walt Disney Inc." to be an ORG. Similarly, if any named entity segment is tagged as a Location, but it is followed or follows a segment tagged as an Organization, we update the labels of the both the Location segment and the Organization segment to be ORG.

Additionally, we observe that many instances of the CoNLL-2003 validation set consist of terse reports of scores of games between sports teams (which are Organization entities), but which also semantically overlap with Location names. For example, the name "Somerset" could refer to a county in England (LOC) or a cricket club (ORG). It is common for a lightly supervised model to confuse the labels to be assigned to such examples. To remedy this, we propose an additional heuristic, which identifies segments labeled as Locations (LOC) and if these segments are followed a score token or at least two integer numbers resembling a score[7], we force their labels to be ORG.

3.4. One Sense Per Discourse

(Amalvy et al., 2023) demonstrated the significance of both local and global document-level context in enhancing the efficacy of pre-trained transformer-based models for NER. In our work, we harness the document-level metadata provided in CoNLL-03 to integrate the "One Sense Per Discourse" (OSPD) principle (Gale et al., 1992) into our neuro-symbolic approach. Primarily conceived for word sense disambiguation, OSPD posits that a term's sense remains consistent when repeatedly used within a cohesive discourse. We operationalize this idea by asserting that if a named entity's predominant classification within a CoNLL discourse leans towards a particular label, then all instances of that entity within the discourse should adopt this dominant label. For instance, should "IBM" appear five times in a document—thrice as an ORG and twice as a LOC—our method dictates that all mentions of "IBM" be labeled as ORG due to its majority occurrence.

3.5. Confidence-Based Rules

In semi-supervised learning, classifier confidence can effectively guide the inclusion of unlabeled data. Nonetheless, contemporary deep neural networks often produce overconfident predictions. To harness the confidence-based heuristics outlined in (Liao and Veeramachaneni, 2009), we adopt the "Smoothed Generalized Cross Entropy" loss from (Zhang and Sabuncu, 2018) & (Dimachkie, 2023), which has been shown to regulate and calibrate model predictions. We then include the following rules in our method: For any segment of tokens classified as an ORG, LOC, or PER with a classifier confidence score $> T$[8], we find other mentions of the same segment within the same CoNLL document and force their label(s) to be the same as the high confidence segment. This is known as the *multi-mention* heuristic. In addition, if the high confidence segment ends in a company suffix, we remove the company suffix and apply the multi-mention property on the remaining segment. We apply the same rule for a high confidence segment that begins with a Person title (from a list of common English honorifics). Furthermore, for each segment ending in a company suffix or starting with a Person prefix, we remove the affix, while retaining the context, to form a new, previously unseen sentence which we then reclassify. For example, suppose we have a sentence in the training data with a PER segment tagged with high confidence as follows: "*The meeting was led by **Ms. Taylor***." Then, removing the Person title would yield the new sentence: "*The meeting was led by **Taylor***." Should the predicted labels of this altered segment in the new sentence differ from those before the affix removal, especially if classified without high confidence, we designate such sentences for inclusion in subset D. This subset is reintroduced to the training data in the subsequent semi-supervised learning iteration, as illustrated in figure 1. We apply these confidence-based heuristics in a sieve i.e. in order of decreasing precision.

3.6. Minimizing The Dependency On A Lexicon

We minimize the dependency of our self-training algorithm on the lexicon chosen by the domain expert by outlining a process that they must follow for picking the lexicon. Using the simple regular expression defined in Section 3.1 (that is able to detect named entity boundaries with high precision), we harvest named entity candidates from the CoNLL-2003 training data in a fully unsupervised

[7] We use regular expressions to detect score tokens, integer patterns, and hyphens respectively.

[8] In our experiments, we empirically set T to 0.9.

manner. These candidates, ranked by their frequency of occurrence in the data, are presented to the domain expert who is tasked to select the most frequent and *unambiguous* ones for each class (i.e. for each class, the lexicon should not contain entities that overlap with another class). By enforcing this criteria of objectivity, we implicitly minimize the chances of multiple domain experts picking vastly different lexicons, thereby minimizing the effects of lexicon variability on our method.

3.7. ELLEN: Integrating Neural And Symbolic Components

In Figure 1, we depict a three-stage framework for amalgamating the heuristics and determining the subset D from unlabeled data U to augment the labeled set L in each semi-supervised learning iteration. This procedure involves three stages: initial burn-in, subsequent intermediate stage, and a concluding burn-out stage. Our selection criterion for D is straightforward: only sentences with entity label updates due to heuristics are considered. This approach aims to expand model knowledge within each cycle while curtailing classic self-training pitfalls. During burn-in, predictions from the Masked Language Modeling Heuristic (MLM) are combined with those from the NER, favoring the MLM due to the NER's initial weakness. Confidence-based heuristics, reliant on the NER's outputs, are deferred. Global rules and OSPD are applied solely on sentences modified by the MLM. As the NER matures through training on MLM outputs, the intermediate phase gives it precedence over the MLM; here, confidence-based heuristics solely target NER predictions, while global rules and OSPD extend to any NER or model-updated sentence. In the burn-out phase, we relax constraints, allowing full self-training. With the model now robust, it gleans any residual data from U for a final training iteration.

4. Experimental Results

4.1. Data & Setup

We evaluate our method on CoNLL-2003 using three different degrees of supervision (see Table 4). We define the "**degree of supervision**" to be the percentage of named entities annotated, relative to the total number of entities present in the data. The first setting is the proposed extremely lightly supervised setting, equivalent to 9.13% in terms of degree of supervision or about "1%" in terms of the number of labeled sentences. We borrow the second "5%" data setting (which corresponds to the first 700 sentences in CoNLL-03's training split) from the current state-of-the-art approaches on semi-supervised NER (Chen et al., 2020; Zheng et al., 2023). However, unlike Chen et al. (2020), which uses `Fairseq` (Ott et al., 2019) for augmenting the unlabeled data with equivalent back-translations from German, we sample 10,000 unlabeled sentences at random[9], without any augmentation. The third setting is the fully supervised setting, where we evaluate the effectiveness of ELLEN against ACE (Wang et al., 2021), the current SOTA method on CoNLL-03, and a DeBERTa V3 (He et al., 2023) classifier[10] finetuned on CoNLL-03.

To summarize, the three different sources of supervision in our experiments, are as follows:

1. **1% data setting:** We use the unambiguous lexicon produced by the domain expert consisting of 10 examples for each of the four CoNLL-03 classes: `MISC`, `ORG`, `LOC` and `PER`.
2. **5% data setting:** In this setting, we also extract an unambiguous lexicon from the entities within the first 700 sentences of the CoNLL-03 training split, adhering to the consistent definition of 'unambiguous' as described in Section 3.6—entities within each class must not overlap with those from other classes. This lexicon, comprising 98 examples for the `MISC` class, 174 for `ORG`, 189 for `LOC`, and 274 for `PER`, is then used for annotating the unlabeled data and for the MLM.
3. **Fully supervised setting:** In this setting, we *do not* use a lexicon for annotating the data since all gold labels are available. However, since the MLM heuristic (section 3.1) requires a lexicon, we extract an unambiguous one just for the MLM, from all of the labeled sentences in CoNLL-03's training split. This lexicon contains 868 examples for the `MISC` class, 2329 for `ORG`, 1245 for `LOC`, and 3598 for `PER` (refer to Appendix C).

4.2. Results Using 1% Labeled Data

As shown in table 5, in the extremely lightly supervised setting, which is more restrictive than typical semi-supervised NER approaches, we find that our method achieves an F1 score of **76.87%** on the CoNLL-2003 test set. The only supervision here comes from a domain expert's lexicon which itself does not use any gold labels from CoNLL-2003. This result indicates our method's real-world effectiveness, where annotations are scarce and lexicons like ours are the only source(s) of supervision.

[9] We do 3 random augmentations for choosing the 10,000 unlabeled sentences. The first 700 sentences are chosen without randomization, to keep the data setting exactly the same as Semi-LADA (Chen et al., 2020) & JointProp (Zheng et al., 2023).

[10] https://huggingface.co/tner/deberta-v3-large-conll2003

Setting	1%	5%	Supervised
Supervision degree	9.13%	28.5%	100%
# Labeled tokens	2569	6971	34043

Table 4: Statistics on degrees of supervision used in this work. 5% (in terms of number of sentences) is a common setting for semi-supervised NER. For the 1% and 5% settings, we calculate the supervision degree based on unambiguous lexicons.

Type	Precision	Recall	F1
Overall	74.63 $\pm 0.33\%$	79.26 $\pm 0.92\%$	76.87 $\pm 0.48\%$
LOC	87.92 $\pm 1.21\%$	78.68 $\pm 4.76\%$	83.04 $\pm 2.36\%$
MISC	56.32 $\pm 1.82\%$	61.00 $\pm 1.25\%$	58.57 $\pm 0.53\%$
ORG	62.29 $\pm 0.39\%$	77.26 $\pm 1.64\%$	68.97 $\pm 0.51\%$
PER	87.98 $\pm 0.92\%$	92.71 $\pm 0.27\%$	90.28 $\pm 0.43\%$

Table 5: Precision/Recall/F1 scores for ELLEN on CoNLL-2003 under the extremely lightly supervised setting. All of our runs are averaged over 3 random seeds. We present entity-level metrics using the official CoNLL-scoring script.

4.3. Results Using 5% Labeled Data

Under the 5% data setting (or 28.5% in terms of "degree of supervision"), we show that our method achieves an F1 score of **84.87**% (Table 6), outperforming more complex methods like PU learning (Zhou et al., 2022), models based on hierarchical latent variables (Chen et al., 2019) & those employing noise strategies (Lakshmi Narayan et al., 2019). We find that it also performs favorably compared to Semi-LADA (Chen et al., 2020) without using any back-translations for the unlabeled data. More importantly, we highlight that our method outperforms PU-Learning approaches whilst using much fewer exemplars per class (the PU-Learning methods of (Peng et al., 2019) & (Zhou et al., 2022) rely on a lexicon that contains "2,000 person names, 748 location names, 353 organization names, and 104 MISC entities"). We include Table 10 (Appendix A), which is directly taken from (Peng et al., 2019), to illustrate this.

Furthermore, we also highlight Figure 4 from GPT-NER (Wang et al., 2023) which shows the performance of ACE (Wang et al., 2021) in a low-resource context. Specifically, when ACE is trained on a 1% subset (in terms of the number of sentences) of CoNLL-03's training data, it's F1 score falls below 20% and below 70% when trained on a 5% subset. Although a fair comparison with our method cannot be made due to differing definitions of "low resource," it is noteworthy that ELLEN attains F1 scores of 76.87% and 84.87% under our equivalent "1%" and "5%" settings respectively, suggesting that our method significantly outperforms ACE in resource-constrained scenarios.

Methods	P	R	F1
VSL-GG-Hier	84.13%	82.64%	83.38%
MT + Noise	83.74%	81.49%	82.60%
Semi-LADA	86.93%	85.74%	86.33%
Jointprop	**89.88%**	**85.98%**	**87.68%**
PU-Learning	85.79%	81.03%	83.34%
ELLEN†	81.88 $\pm 1.18\%$	88.01 $\pm 0.19\%$	84.87 $\pm 0.62\%$

Table 6: Performance on CoNLL 2003 with 5% labeled data. It should be noted that JointProp (Zheng et al., 2023) is a multi-task learning framework. All of our runs are averaged over 3 random seeds.

4.4. Zero-Shot Evaluation

We apply our extremely lightly supervised ("1%") method in a zero-shot manner on WNUT-17, a dataset from the social media domain, characterized by noisy text. That is, using the model that was trained on CoNLL-03 with only a lexicon of 10 samples per class (provided by the domain expert), we proceed to evaluate this model on the WNUT-17 test dataset. After aligning the predictions of each model with the label space of CoNLL-03 (see Appendix B for details), we observe that ELLEN achieves comparable zero-shot performance to GPT-3.5 and GPT-4, and also achieves relatively strong zero-shot performance against a *fully supervised* model[11] from the the T-NER library (Ushio and Camacho-Collados, 2021) that was actually trained on WNUT-17 gold data (see Table 7). This result is exciting because it indicates the potential for our method to be used across domains, given the relatively small size of our model and its extremely light supervision.

4.5. Results Using Full Supervision

Lastly, we show that our method can also be adapted to a fully-supervised setting. Table 8 shows that we obtain a respectable F1 score of

[11]We use the RoBERTa large model, available here: https://huggingface.co/tner/roberta-large-wnut2017

Method	LOC	MISC	ORG	PER	AVG
T-NER	**64.21%**	**42.04%**	**42.98%**	**66.11%**	**55.11%**
GPT-3.5	49.17%	8.06%	29.71%	59.84%	39.96%
GPT-4	58.70%	25.40%	38.05%	56.87%	43.72%
ELLEN†	44.82 ±3.84%	6.21 ±1.25%	26.49 ±5.01%	67.00 ±3.54%	41.56 ±0.92%

Table 7: Comparing F1 scores: ELLEN, GPT-3.5, and GPT-4 are evaluated in zero-shot mode against T-NER's fully supervised model on the WNUT-17 test set, after label alignment with CoNLL-03 († indicates our framework). For ELLEN, we report the average zero-shot score of 3 different random initializations and training runs of the models under extremely light supervision.

90.98% relative to ACE (**94.6%**) and a standard supervised classifier (**92.2%**). We note that, while our neuro-symbolic approach is effective in low resource settings, when full supervision is available, other methods may outperform ours. This is primarily due to the noise introduced by the various heuristics we propose in Section 3 (MLM, One Sense Per Discourse, confidence-based rules), which may erroneously annotate O/Outside entities as belonging to a non-O class, leading to our model being iteratively retrained on some noisy data (as shown in Figure 1).

Model	F1
ELLEN	90.98 ±0.54%
DeBERTa V3	92.2%
ACE (Wang et al., 2021)	**94.6%**

Table 8: Performance of ELLEN on CoNLL-2003 test when using all available annotations from the CoNLL-2003 training data (fully supervised setting). For ELLEN, we report an average of training runs over 3 random seeds.

4.6. Error Analysis And Ablation Experiment

Focusing on the extremely lightly supervised setting for CoNLL-03, we observed that over 30% of model errors on validation data involve confusing ORG and LOC entities. These errors can be attributed to a combination of factors: a) a bias in CoNLL-03's validation and test data towards sporting events not sufficiently reflected in the training data, b) the inadequacy of a global rule (refer to Section 3.3) to differentiate between sports teams (ORG) and locations (LOC) in nuanced contexts, e.g., both 'YORKSHIRE' and 'HEADINGLEY' would be labeled as ORG's in the sentence: "YORKSHIRE AT HEADINGLEY" even though 'HEADINGLEY' is a LOC. c) errors arising from noisy POS tags and incorrectly identified entity spans, e.g., "Dhaka Stock Exchange" would be identified as two separate entities "Dhaka" and "Stock Exchange" by the regular expression, leading to incorrect labels by the MLM Heuristic during training; and d) confusion between ORG and MISC classes, partly because the MISC class lexicon primarily includes nationalities, which does not fully represent its broader scope (e.g., events, products, works of art).

In an ablation on the CoNLL-2003 dev set (Table 9), we found the MLM to be the most impactful component. This was followed by dynamic window filtering, which allows our method to achieve a **64.7%** F1 score on its own. Importantly, reintegrating other components—OSPD, confidence-based heuristics, and global rules—each further enhances performance, underscoring their collective contribution to the method's effectiveness.

Ablations	P	R	F1
Full system	71.31 ±2.4%	75.90 ±0.9%	73.52 ±1.6%
MLM	61.96 ±1.1%	74.00 ±1.0%	67.40 ±0.2%
CR, GR, OSPD	70.30 ±3.5%	74.36 ±0.8%	72.23 ±2.1%
MLM, CR, GR, OSPD	59.20 ±1.8%	71.30 ±1.0%	64.70 ±1.3%

Table 9: Ablation of major components in our system, measured by P/R/F1 on CoNLL-03 validation data. **'CR'** refers to "Confidence-Based Rules." **'GR'** refers to "Global Rules." **'OSPD'** refers to "One Sense Per Discourse". **'MLM'** refers to the "Masked Language Model".

5. Conclusion

In this paper, we present a framework that harmoniously blends linguistics and deep learning to overcome the paucity of labeled data for NER, requiring significantly less supervision than previous methods. Real-world entity extraction is often hindered by the lack of annotated data, especially in low-resource domains. While LLMs offer potential remedies, they are not without limitations. Our solution, ELLEN, introduces an efficient, encoder-only method that enables the assembly of an NER system in as little as "half a day", requiring only a single expert-provided lexicon. We show ELLEN's strong performance in the extremely low resource setting, showing that it scales well under varying supervision levels, while also outperforming other, more complex approaches.

Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency (DARPA) under the Habitus program and by the National Science Foundation (NSF) under grant #2006583. Mihai Surdeanu declares a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

Limitations

Our proposed method, while showcasing promising results in settings of lightly supervised named entity recognition (NER), faces certain limitations that warrant discussion. Primarily, our evaluation was conducted only on two flat NER datasets. Adapting our method across a broader spectrum of datasets, especially those that may feature more complex, fine-grained, or nested entity structures, needs further exploration. Consequently, our current approach does not explicitly address the challenges associated with more intricate NER tasks, such as nested, fine-grained, hierarchical or intersectional NER, which require the identification of entities within entities or the recognition of novel entity types beyond traditional categories.

Some of the rules employed by our method are domain and language-specific, which could limit their wider applicability. However, we also highlight that four out of the eleven total rules in our method are domain and language-independent (assuming the existence of a language model for that domain/language). These include the Masked Language Model (MLM), a heuristic for "free" supervision from exemplars (which can come from any domain or language), Dynamic Window Filtering, which assumes that POS tags are available for a given language/domain and that the language distinguishes between common nouns and proper nouns, One Sense Per Discourse (OSPD) which simply propagates the majority label within a document, and the basic multi-mention heuristic for label propagation which only uses classifier confidence scores. MLM and dynamic window filtering, both language and domain-independent, are the two components that contribute the most to the performance of our NER method (as shown in Table 9).

Certain rules, such as those dependent on company suffixes and person honorifics, are not domain-independent but are transferable across languages with the adaptation of language-specific affixes. This adaptability suggests a pathway to applying our method to new languages, provided a list of relevant suffixes and honorifics is used. However, there may be challenges in directly applying some of these rules to languages which use vastly different conventions for naming entities. Nevertheless, in presenting our findings, we have not claimed our method to be universally applicable across all languages and domains. Instead, we aimed to demonstrate how the integration of linguistic insights with neural networks can mitigate the scarcity of labeled data in NER. Our framework, which harmoniously blends these elements, points to a significant step forward, while highlighting the necessity for further research to extend its applicability to more diverse and complex NER scenarios.

Ethical Considerations

This work utilizes two public, commonly-used datasets for Named Entity Recognition (NER). One dataset, WNUT-17, derived from the social media domain, mainly consists of user-generated comments, of which a very small portion may include language that some might find inappropriate or offensive. Furthermore, our approach incorporates open-source pre-trained language models. Thus, any biases inherent in these models due to their pre-training data would also apply to our work. Regarding the selection of named entity seeds (see Section 3.6), while efforts were made to minimize subjectivity in the creation of the lexicon, it is theoretically possible for the proposed method to be used to intentionally introduce biases into an NER model. However, we believe that, apart from the potential issues already mentioned, this work does not raise any significant ethical concerns.

6. Bibliographical References

Arthur Amalvy, Vincent Labatut, and Richard Dufour. 2023. The role of global and local context in named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 714–722, Toronto, Canada. Association for Computational Linguistics.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Jatin Arora and Youngja Park. 2023. Split-NER: Named entity recognition via two question-answering-based classifications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short*

Papers), pages 416–426, Toronto, Canada. Association for Computational Linguistics.

Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: Prompting for named entity recognition.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Haiyan Chen, Shuwei Yuan, and Xiang Zhang. 2022. Rose-ner: Robust semi-supervised named entity recognition on insufficient labeled data. In *Proceedings of the 10th International Joint Conference on Knowledge Graphs*, IJCKG '21, page 38–44, New York, NY, USA. Association for Computing Machinery.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020. Local additivity based data augmentation for semi-supervised NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1241–1251, Online. Association for Computational Linguistics.

Jiawei Chen, Yaojie Lu, Hongyu Lin, Jie Lou, Wei Jia, Dai Dai, Hua Wu, Boxi Cao, Xianpei Han, and Le Sun. 2023. Learning in-context learning for named entity recognition.

Mingda Chen, Qingming Tang, Karen Livescu, and Kevin Gimpel. 2019. Variational sequential labelers for semi-supervised learning.

Kevin Clark, Thang Luong, Christopher D. Manning, and Quoc V. Le. 2018. Semi-supervised sequence modeling with cross-view training.

Chady Dimachkie. 2023. Cross-entropy is all you need... or is it? Medium. Accessed: 2023-10-17.

William A Gale, Kenneth Church, and David Yarowsky. 1992. One sense per discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *CoRR*, abs/2308.08493.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Carina Kauf and Anna A. Ivanova. 2023. A better way to do masked language model scoring. *ArXiv*, abs/2305.10588.

Pooja Lakshmi Narayan, Ajay Nagesh, and Mihai Surdeanu. 2019. Exploration of noise strategies in semi-supervised named entity classification. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 186–191, Minneapolis, Minnesota. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Yangming Li, Lemao Liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition.

Wenhui Liao and Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, page 58–65, USA. Association for Computational Linguistics.

Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. 2019a. Towards improving neural named entity recognition with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5301–5307, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tong Luo, Kurt Kramer, Dmitry B. Goldgof, Lawrence O. Hall, Scott Samson, Andrew Remsen, and Thomas Hopkins. 2005. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, 6(20):589–613.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ajay Nagesh and Mihai Surdeanu. 2018. Keep your bearings: Lightly-supervised information extraction with ladder networks that avoids semantic drift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 352–358, New Orleans, Louisiana. Association for Computational Linguistics.

Deepak Narayanan, Keshav Santhanam, Peter Henderson, Rishi Bommasani, Tony Lee, and Percy Liang. 2023. Cheaply evaluating inference efficiency metrics for autoregressive transformer apis.

OpenAI. 2023a. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2023b. Introducing chatgpt. Accessed: 10-20-2023.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy. Association for Computational Linguistics.

Mrinmaya Sachan and Eric Xing. 2016. Easy questions first? a case study on curriculum learning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 453–463.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023. Lm contamination index. Accessed: 2023-10-12.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. Prompt-ner: Prompt locating and typing for named entity recognition.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.

Robert Vacareanu, Enrique Noriega-Atala, Gus Hahn-Powell, Marco A. Valenzuela-Escárcega, and Mihai Surdeanu. 2024. Active learning design choices for NER with transformers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italy. European Language Resources Association.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated Concatenation of Embeddings for Structured Prediction. In *the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (**ACL-IJCNLP 2021**)*. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot ner with chatgpt.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels.

Yandan Zheng, Anran Hao, and Anh Tuan Luu. 2023. Jointprop: Joint semi-supervised learning for entity and relation extraction with heterogeneous graph-based propagation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14541–14555, Toronto, Canada. Association for Computational Linguistics.

Kang Zhou, Yuepei Li, and Qi Li. 2022. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7198–7211, Dublin, Ireland. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition.

7. Language Resource References

Derczynski, Leon and Nichols, Eric and van Erp, Marieke and Limsopatham, Nut. 2017. *Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition*. Association for Computational Linguistics.

Tjong Kim Sang, Erik F. and De Meulder, Fien. 2003. *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*.

A. Statistics of Labeling With PU Learning Lexicon

Type	# of 1.w.	Precision	Recall
PER	2,507	89.26%	17.38%
LOC	4,384	85.07%	50.03%
ORG	3,198	86.17%	29.45%
MISC	1,464	92.13%	30.59%

Table 10: Data labeling results using the lexicon used by PU Learning methods: the number of labeled words (# of l.w.), the word-level precision (# of true labeled words/# of total labeled words) and recall, on CoNLL-2003.

Table 10, which is directly taken from the work of (Peng et al., 2019), illustrates the data labeling statistics of the large lexicons (sourced from external dictionaries) used by PU-learning methods. This is in stark contrast to our method, where the lexicon contains only 10 exemplars per class in the '1%' data setting or at most, a few hundred for the `PER` class in the '5%' data setting.

B. WNUT-17: Zero-Shot Evaluation

To allow our model, which was trained on CoNLL-2003, to be fairly compared in a zero-shot setting against the fully supervised `roberta-large` model from the T-NER (Ushio and Camacho-Collados, 2021) library (which is trained on WNUT-17 gold data), we mapped the generated labels from the fully supervised model onto the label space of CoNLL-2003. We aligned the classes from WNUT-17 with CoNLL classes (`ORG, LOC, PER, MISC`) based on semantic overlap. We aligned the 'products' and 'creative-work' classes with the MISC class from CoNLL. This is because the MISC class from CoNLL also contains many product names and 'works of art,' e.g., "Ain't No Telling" by Jimi Hendrix.

We aligned the 'group' class from WNUT-17 with 'ORG' from CoNLL because many 'group' names in the WNUT-17 test data have a semantic overlap with organizations, e.g., 'Nirvana", 'San Diego Padres.' Based upon our inspection of the data, the 'group' class also includes entities like musical bands, sports teams, non-profit organizations, political groups, etc. Such entities fit well within the typical CoNLL understanding of an 'organization.' For the remaining classes of WNUT-17 (`corporation, location, person`), we mapped them directly to their corresponding CoNLL-03 equivalents. We also applied this mapping to the zero-shot predictions of GPT-3.5 and GPT-4, to allow all models to be fairly compared against each other. We evaluated all models shown in Table 7 on the full test set (1287 samples) of WNUT-17.

We accessed GPT-3.5 and GPT-4 through the Azure OpenAI service, using the `gpt-35-turbo-0613` and `gpt-4-0613` models with `temperature=0` for deterministic results. We borrow the prompt format from the vanilla zero-shot prompt used by (Xie et al., 2023) (shown in the figure below). In the zero-shot evaluation with GPT-3.5 and GPT-4, we observed issues similar to those observed by (Wang et al., 2023), i.e., both LLMs often fail to match the output length with the input sentence's token count in sequence labeling tasks like NER, a challenge amplified in longer sentences. This is documented in Table 11, distinguishing "Misalignment errors"—the discrepancy in the number of LLM generated NER tags versus sentence tokens—and "Parsing errors," where the LLM generation does not form a valid sequence of NER labels and hence, cannot be parsed, with GPT-3.5 showing more pronounced issues.

To mitigate these alignment problems, we used a simple approach:

1. For outputs with fewer NER tags than input tokens, we padded the sequence on the right with '`O`' tags to equalize the lengths.

2. For outputs with excess NER tags, we truncated the surplus from the right to match the input token sequence length.

We then evaluated the aligned and corrected predictions of both LLM's on the WNUT-17 test set using the official CoNLL-scoring script (results reported in Table 7).

Model	Misalignment Errors	Parsing Errors
GPT-3.5	426	80
GPT-4	195	44

Table 11: Comparison of error counts between GPT-3.5 and GPT-4.

Prompt Used For Zero-Shot Evaluation of GPT-3.5/4

Given entity label set: ['B-PER', 'I-PER', 'B-ORG', 'I-ORG', 'B-LOC', 'I-LOC', 'B-MISC', 'I-MISC', 'O']
Based on the given entity label set, please recognize the named entities for each token in the given text, and return the answer as a list of named entity tags.
Text: {input text}

Answer: {ChatGPT response}

C. Masked Language Model (MLM): Inverse Breaking Ties

In order to obtain more robust annotations from the Masked Language Model (MLM), we only consider the entity span x_i labeled by the MLM where the difference between the score of the class that is predicted with the highest probability and the score of the class that is predicted with the second highest probability is greater than some threshold t_{class}:

$$x_i \mid P(y_i = l_1|x_i) - P(y_i = l_2|x_i) > t_{class}$$

Here l_1 is the most likely class label and l_2 is the second most likely class label, according to the MLM. This is motivated by the *Breaking Ties* active learning method of Scheffer et al. (2001); Luo et al. (2005), which aims to select token samples where the difference between the top two predictions is the smallest, in order to increase the likelihood of confident classifications. However, for the MLM, we adopt the *inverse* of breaking ties, where we maximize the difference between the top two predictions, based on a threshold. We use different thresholds for each class as shown in Table 12. We empirically observe that we obtain a slightly higher F1 score with the MLM as an unsupervised NER on the CoNLL-03 development set when using different thresholds for each class instead of a single threshold value for all classes.

In our experiments, we also empirically observe that the MLM tends to produce more robust probabilities when the lexicon entities filling the `[MASK]` slot(s) are segmented into fewer subwords by the model's tokenizer. This is supported by the findings of Kauf and Ivanova (2023), who observe that methods that estimate the psuedo-log-likelihood of a sentence yield inflated scores for out-of-vocabulary words. Hence we employ an additional heuristic where we filter the lexicon entities to only single subword entities. We believe that better methods for estimating and aggregating probabilities for sentences that contain out-of-vocabulary words can be explored in future work. Kauf and Ivanova (2023) introduce one such method, which has been shown to address the issue of attributing uneven likelihoods to multi-token words. Specifically, it proves beneficial to mask not only the current token but also all subsequent tokens that are part of the same word.

Furthermore, given the large size of the lexicons extracted for the MLM in both the "5%" and fully supervised setting, we filter our lexicon to keep only the top 20 entities for each class, sorted by their frequency of occurrence in the training data.

Class	Threshold
ORG	0.28
PER	0.2
LOC	0.1
MISC	0.05

Table 12: Per class thresholds used by the Masked Language Model (MLM) to implement "inverse breaking ties".

D. Hyperparameters and Hardware

Instead of the regular cross-entropy loss, we use a Generalized Cross Entropy Loss function with label smoothing (Dimachkie, 2023) for training our models, which offers a better trade-off between the noise-robustness of mean absolute error and the noise sensitivity of cross entropy loss. This trade-off can be controlled by a hyperparameter q. We experiment with multiple settings where we vary q, along with the learning rate, the dynamic window size, the number of burn-in, intermediate and burn-out stages and the total number of self-training iterations. This search involved under 20 runs, based on the development partition of CoNLL-2003. We use a dynamic window size of 5, a batch size of 16 for training, a learning rate of $1e{-}5$, and a confidence-threshold of 0.9 across all of our data settings. We show the other hyperparameters in Table 13. All experiments were carried out on a system with 2 Nvidia RTX 3090 GPUs.

Setting	Burn-in stages	Intermediate stages	Burn-out stages	Noise-level (q)	Self-training iterations	Label Smoothing
1% data	1	2	1	0.9	4	0.1
5% data	1	2	0	0.7	3	0.1
100% data	1	1	0	0.7	2	0.2

Table 13: The Hyperparameters we use for training ELLEN under various supervision settings.