Large-Scale Label Interpretation Learning for Few-Shot Named Entity Recognition

Anonymous ACL submission

Abstract

Few-shot named entity recognition (NER) detects named entities within text using only a few annotated examples. One promising line of research is to leverage natural language descriptions of each entity type: the common label PER might, for example, be verbalized as "person entity." In an initial *label interpretation learning* phase, the model learns to interpret such verbalized descriptions of entity types. In a subsequent *few-shot tagset extension* phase, this model is then given a description of a previously unseen entity type (such as "music album") and optionally a few training examples to perform few-shot NER for this type. In this paper, we systematically explore the impact of massively scaling up the number and granularity of entity types used for label interpretation learning. To this end, we leverage WikiData to create a dataset with orders of magnitude of more distinct entity types and descriptions as currently used datasets. We find that this increased signal yields strong results in zero- and few-shot NER in in-domain, cross-domain, and even cross-lingual settings (e.g. increasing F1 ↑14.7 pp. on FewNERD and ↑9.0 pp. on Chinese OntoNotes). Our findings indicate significant potential for improving few-shot NER through heuristical data-based optimization.

1 Introduction

Few-shot named entity recognition (NER) refers to identifying and classifying named entities within text by learning from a few annotated examples. A widely adopted strategy in few-shot NER employs transfer learning with pre-trained language models (PLMs) to interpret labels based on their semantic meaning (Yang and Katiyar, 2020; de Lichy et al., 2021; Das et al., 2022; Ma et al., 2022a,b,c; Chen et al., 2023; Zhang et al., 2023). The main idea is that such models learn to interpret a natural language description of an entity type for use in a word-level decoder. They learn in two phases:

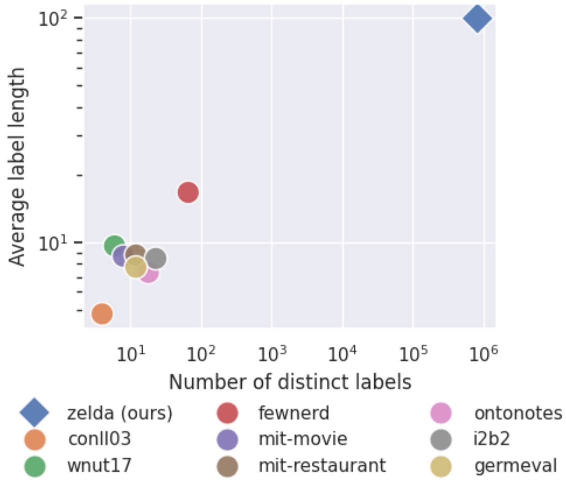

Figure 1: Few-shot NER requires an initial label interpretation learning phase using the entity types of a source dataset. We propose learning from orders of magnitude more distinct types and more expressive label semantics than current NER datasets by using existing entity linking datasets annotated with WikiData information.

1. a *label interpretation learning* phase on an NER-annotated dataset with a set of entity types and their verbalizations. For instance, the common label PER might be verbalized as "person entity." In this phase, the model learns to associate entity type verbalizations with matching NER annotations.

2. a *few-shot tagset extension phase* in which the model is expanded to previously unseen entity types using only a new verbalization and optionally a few example annotations. For instance, to extend the model to recognize the names of music albums, one would only need to provide a verbalization ("music album") and a few examples.

Data limitations. However, as Figure 1 indicates, prior studies used only very limited numbers of distinct entity types for label interpretation learning. This is an artifact of relying on common NER datasets such as CoNLL-03 (Tjong Kim Sang and De Meulder, 2003), OntoNotes (Pradhan et al., 2012), WNUT-17 (Derczynski et al., 2017), or FewNERD (Ding et al., 2021), which only contain a small number of distinct entity types (between 4 and 66 types). Furthermore, the majority of their entity types have a simple semantic definition, such as "person," "location," or "organization," and occur across several datasets. We hypothesize that these limitations overly constrain the semantic signal that is observed during label interpretation learning, thus constituting a main limiting factor to few-shot NER.

Contributions. With this paper, we introduce LITSET (label interpretation learning by scaling entity typing) and systematically investigate the intuition that increasing the number of distinct entity types and their descriptive granularity in label interpretation learning improves few-shot NER capability. To this end, we heuristically create a dataset with orders of magnitude more distinct entity types than commonly employed (see Figure 1) and use it for extensive experimentation. In more detail, our contributions are:

- We present experiments to validate our hypothesis on the largest existing NER dataset (FewNERD). We find that few-shot performance increases with label interpretation learning on more distinct entity types and more expressive descriptions (cf. Section 2).
- To massively scale up label interpretation learning, we present an approach for deriving a dataset with orders of magnitude more granular entity type annotations. Our approach leverages an existing entity linking dataset and enriches it with type descriptions from WikiData (Vrandečić and Krötzsch, 2014) (cf. Section 3).
- We comprehensibly evaluate label interpretation learning on our derived corpus against classical setups for zero- and few-shot NER in in-domain, cross-domain, and cross-lingual settings (cf. Section 4).

We find that label interpretation learning on our heuristically derived corpus matches and, in many cases, significantly outperforms strong baselines. Our findings indicate significant potential for improving few-shot NER through heuristical data-based optimization.

To enable the research community to reproduce and leverage this work, we release the generated dataset and source code under the Apache 2 license at: *(inserted after review)*

2 Validation Experiment for Impact of Entity Types and Label Descriptions

We first conduct an experiment to validate the intuition that a richer training signal for label interpretation learning positively impacts few-shot NER. To this end, we create a set of training datasets for label interpretation learning that each contain the same number of entities but vary in the number of distinct entity types and their label verbalization. We then compare the few-shot NER ability of models trained on each of these datasets.

2.1 Experimental Setup

Definitions. To evaluate few-shot NER, an existing dataset $\mathcal{D}$ is split based on its labels $\mathcal{L}$: the label interpretation training split $\mathcal{D}^{LIT}$ and a few-shot fine-tuning split $\mathcal{D}^{FS}$. The corresponding labels of each split $\mathcal{L}^{LIT}$ and $\mathcal{L}^{FS}$ are set such that $\mathcal{L}^{LIT} \cup \mathcal{L}^{FS} = \mathcal{L}$ and $\mathcal{L}^{LIT} \cap \mathcal{L}^{FS} = \emptyset$.

Dataset. We use FewNERD in our experiment since it is the largest existing dataset w.r.t. the number of distinct entity types (66 types). We set the labels of D^{LIT} to be the 50 most occurring entity types and the labels of D^{FS} to be the 16 least occurring. We perform an analysis along two dimensions:

- To measure the impact of increasing the number of distinct entity types in label interpretation learning, we create 5 versions of the training data containing 3, 5, 10, 30, and all 50 labels, respectively. Importantly, all 5 versions are of the same size and contain the same number of labeled entities (10k).
- To measure the impact of richer verbalizations, we define 3 different labels semantics: *(1)* a "cryptic" unique, random 2-character label, *(2)* a "short" description as regularly used according to research and *(3)* a "long" description with examples (cf. Appendix A).

To exclude the respective labels from each split, we follow prior work and mask labels $\mathcal{L}^{LIT}$ in $\mathcal{D}^{FS}$ and $\mathcal{L}^{FS}$ in $\mathcal{D}^{LIT}$ with the `O`-token (meaning no named entity).

Few-shot model. We employ the frequently used bi-encoder architecture (Blevins and Zettlemoyer,

Figure 2: F1 scores for few-shot NER tagset extension depending on how many distinct entity types were seen in label interpretation learning (columns), and how label types were verbalized (rows). We report F1 scores averaged over five seeds. We observe that (*1*) more distinct labels during label interpretation training and (*2*) more semantically expressive labels improve few-shot NER.

2020; Ma et al., 2022a; Zhang et al., 2023) with two `bert-base-uncased` transformers as our backbone architecture. For few-shot tagset extension, we sample a support set $\mathcal{S}$ by k-shot downsampling $\mathcal{D}^{FS}$. The support set $\mathcal{S}$ contains each label from $\mathcal{L}^{FS}$ exactly k times. We sample three different support sets using different seeds and report the averaged micro-F1 scores over these iterations.

2.2 Results

Figure 2 shows the results of tagset extension when performing label interpretation learning on corpora with different numbers of labels (columns) and different verbalization methods (rows). For each label interpretation learning, we report the average F1-score for tagset extension for 1-shot, 5-shot, and 10-shot learning, respectively.

Improved generalization with more types. We observe that the number of distinct labels seen during label interpretation training increases the generalization in few-shot settings independent of the label semantics used. We find improvements from +3.0 F1 (cf. $L = 3$ vs. $L = 50$, label semantic: cryptic) up to 8.7 F1 (cf. $L = 3$ vs. $L = 50$, label semantic: short) on average in pp.

More expressive descriptions helpful. We also find that increasing the expressiveness of label verbalizations strongly improves the few-shot performance. This observation is independent of the number of labels seen in label interpretation learning, such that we find improvements ranging from +16.8 F1 (cf. label semantics: simple vs. long, with $L = 3$) up to 22.0 F1 (cf. label semantics: simple vs. long, with $L = 50$) on average in pp.

These observations support our intuition that a richer training signal in label interpretation learning improves few-shot NER performance.

3 Large-Scale Label Interpretation Learning

As our validation experiment found a positive impact of increasing the number and expressivity of entity types, we now aim to scale the signal for label interpretation learning to orders of magnitude more entity types. To this end, we heuristically derive an NER-annotated dataset we call LITSET using entity disambiguation and WikiData (Section 3.1). We also present a small modification to the bi-encoder network to handle a very large space of entity types (Section 3.2).

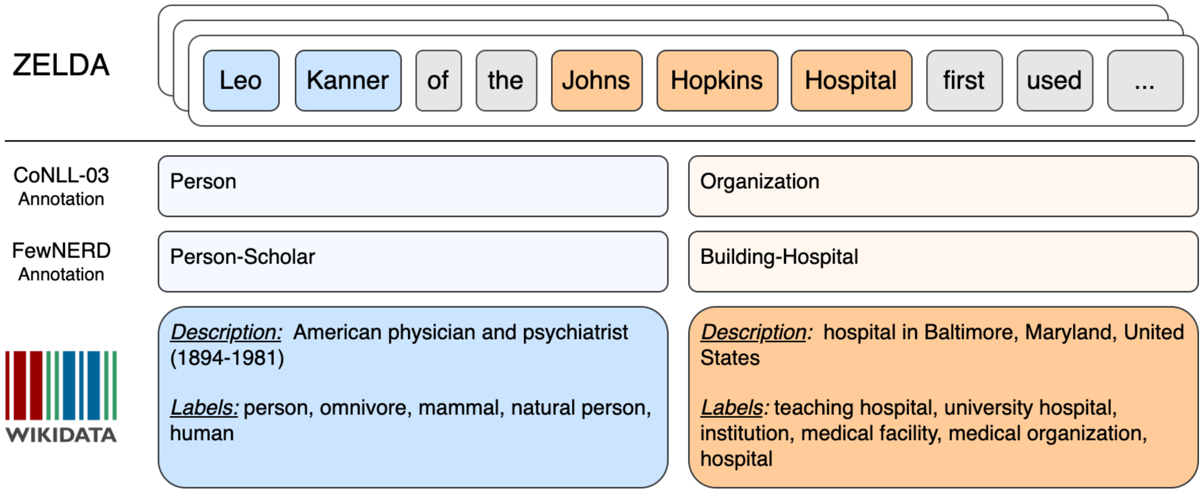

Figure 3: An example annotation of a sentence in ZELDA. WikiData can provide distinct descriptions and labels about the respective entity, whereas the annotations, compared to existing datasets, would be less informative if not misleading.

Dataset	Label length	# Distinct types
CoNLL-03	9.8 ± 2.9	4
WNUT17	8.3 ± 2.8	6
OntoNotes	9.8 ± 8.5	18
FewNERD	17.3 ± 7.6	66
LitSet	99.8 ± 45.4	~817k

Table 1: Average label description length (in characters) and distinct entity types of NER datasets. Label length and distinct entity types for LitSet refers to all annotations as indicated in Figure 3

3.1 LitSet Dataset

The task of entity disambiguation is closely related to NER. Here, an already detected entity is disambiguated by linking it to an existing knowledge base such as Wikipedia or WikiData. Existing training and evaluation datasets for entity disambiguation thus contain named entities marked with links to entries in the WikiData knowledge base.

One advantage of WikiData is that it contains fine-grained labels and free-form text descriptions of entities in the knowledge base. For instance, the entity "John Hopkins Hospital" (see Figure 3) has the free-form description "hospital in Baltimore, Maryland" and belongs to the classes "teaching hospital", "university hospital", and many others. As the Figure shows, these labels are significantly more fine-grained than CoNLL-03 and even FewNERD entity types which simply classify it as an "organization" or a "hospital" respectively.

Deriving LitSet. In our approach, we leverage these classes and descriptions as type annotations. As base entity disambiguation dataset, we use the recently released ZELDA (Milich and Akbik, 2023) benchmark as it represents a broad range of topics, making it a suitable dataset for the general domain. For each linked entity in the dataset, we retrieve the types and descriptions from WikiData and use them as NER annotations.

However, as Figure 3 illustrates, each linked entity belongs to multiple WikiData classes and has a potentially long description. For this reason, we subsample the annotations to bring their length more in line with standard NER datasets. Specifically, for each entity x_i, we uniformly sample whether we annotate it with either the `description` attribute or the `labels` attribute (cf. Figure 3). When utilizing the `labels` attribute, we randomly select the number of tags following a geometric distribution with $p = .5$. Subsequently, we uniformly sample tags from the `label` attribute until the number of tags is reached. Lastly, we concatenate the selected tags for final annotation.

3.2 Backbone Architecture

We conduct our experiments based on the widely adopted bi-encoder model due to its simplicity. The model utilizes two separate transformers to encode tokens and labels, respectively. The first transformer generates embeddings $e_t \in \mathbb{R}^{N \times H}$ for all tokens, where N represents the number of tokens and H denotes the hidden size of the model. The second obtains the `[CLS]`-token embeddings e_l for the labels, which are converted into natural language. We employ cross-entropy loss and derive final predictions with

$$\hat{y} = \arg\max softmax(e_t \cdot e_l)$$

However, training a model, including the bi-encoder, with a wide array of distinct classes is non-trivial. With $\mathcal{L}$ denoting the set of labels, the shape of label representations is $e_l \in \mathbb{R}^{|\mathcal{L}| \times H}$. Given that $|\mathcal{L}| \approx 10^6$ (cf. Figure 1), we aim to circumvent the resulting matrix multiplication for two reasons: (*1*) potential computational limitations and (*2*) optimization difficulty. To alleviate these issues, we restrict our consideration to labels present in the current batch $\mathcal{L}_b$ with $|\mathcal{L}_b| \ll |\mathcal{L}|$ for loss calculation.

While the resulting dataset has the potential to be applied to various few-shot NER methods if the aforementioned issues are addressed, we leave this exploration to future research.

4 Experiments

We evaluate the impact of label interpretation training in various tagset extension settings. Throughout all experiments, we compare label interpretation learning on LITSET with training on different baseline datasets. Specifically, we conduct the following experiments:

1. *In-domain transfer*: Identical domain in label interpretation learning and few-shot fine-tuning (cf. Section 4.1).

2. *Cross-domain transfer*: Different domain in label interpretation learning and few-shot fine-tuning (cf. Section 4.2).

3. *Cross-lingual transfer*: Identical domain in label interpretation learning and few-shot fine-tuning, but languages differ between both phases (cf. Section 4.3).

Further, we support our experiments by analyzing the impact of different label semantics used between label interpretation learning and few-shot fine-tuning (cf. Section 4.1). At last, we refer to our ablation experiments on the impact of different transformer models as label encoder and negative sampling (cf. Appendices D and E).

4.1 Experiment 1: In-Domain Transfer

This experiment replicates the most common evaluation setup for few-shot tagset extension, where both $\mathcal{D}^{LIT}$ and $\mathcal{D}^{FS}$ are sourced from the same NER dataset. Our baseline is the default approach of label interpretation learning on $\mathcal{D}^{LIT}$, which is "in-domain" since it shares the same textual domain and entity granularity as the evaluation data. We compare this baseline against label interpretation learning on LITSET.

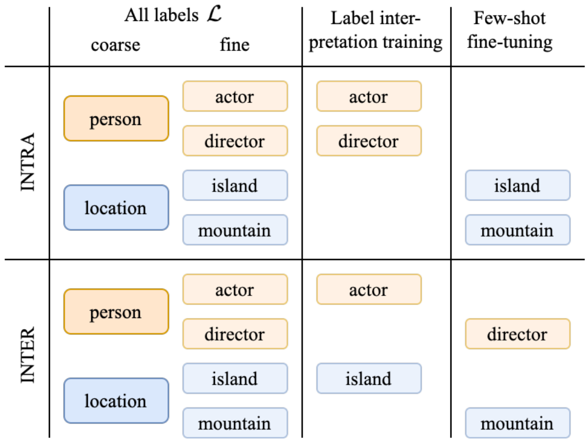

Figure 4: Exemplary illustration on the INTRA and INTER settings of FewNERD experiments.

4.1.1 Experimental Setup

We use OntoNotes and FewNERD in our experiments, as they have important properties: OntoNotes covers different domains and languages such that we can measure the transferability of our approach. FewNERD comes with two annotation types: coarse labels $\mathcal{L}^c$ (8 classes) and fine labels $\mathcal{L}^f$ (66 classes). $\mathcal{L}^f$ are subclasses of the $\mathcal{L}^c$ such that the entity mentions of both annotations are identical, only their surface form differs. Thus, we can evaluate our dataset against FewNERD in two ways: *(1)* the INTRA setting in which we split the labels based on coarse annotations, and *(2)* in which we split based on the fine annotations (cf. Figure 4).

We split each dataset into two equally sized label sets. To reduce the impact of randomness, the random split is repeated three times. We then perform few-shot fine-tuning runs with three different seeds for each random split.

Comparison with LITSET. To focus solely on understanding the impact of scaling entity types without the influence of increased entity detection, we downsample LITSET to match the number of entity mentions in each baseline dataset. Further, to make a fair comparison, we remove labels from our approach that match those in the baseline labels $\mathcal{L}^{FS}$ and mask them with the O-token. However,

Evaluation data $\mathcal{D}^{FS}$ for tagset extension from:	Label interpretation learning data $\mathcal{D}^{LIT}$ from:	0-shot	1-shot	5-shot	10-shot	Avg.
FewNERD$_{\text{INTRA}}$	LitSet	3.2 ± 1.0	$\mathbf{30.7} \pm 5.3$	$\mathbf{51.9} \pm 5.2$	$\mathbf{57.9} \pm 6.2$	$\mathbf{35.9}$
	w/ all labels	0.9 ± 0.4	$\underline{20.1} \pm 5.0$	$\underline{47.7} \pm 6.0$	$\underline{54.1} \pm 5.9$	$\underline{30.7}$
	w/ labels only	$\underline{3.7} \pm 0.5$	14.3 ± 8.3	29.6 ± 7.0	37.5 ± 6.1	21.3
	w/ description only	1.0 ± 0.3	19.8 ± 8.8	37.5 ± 7.9	46.2 ± 5.9	26.1
	FewNERD$_{\text{INTRA}}$ *(Baseline)*	$\mathbf{5.8} \pm 0.4$	8.9 ± 4.3	31.4 ± 9.2	38.4 ± 7.5	21.1
OntoNotes	LitSet	$\mathbf{8.7} \pm 1.7$	$\mathbf{21.9} \pm 8.4$	$\mathbf{40.1} \pm 7.2$	$\underline{48.4} \pm 6.2$	$\mathbf{29.5}$
	w/ all labels	3.5 ± 1.3	$\underline{20.0} \pm 9.5$	$\underline{38.4} \pm 8.3$	46.5 ± 6.3	$\underline{27.1}$
	w/ labels only	0.1 ± 0.1	14.3 ± 8.3	29.6 ± 6.9	37.5 ± 6.1	20.4
	w/ description only	$\underline{4.2} \pm 1.3$	19.8 ± 8.8	37.5 ± 7.9	46.2 ± 5.9	26.9
	OntoNotes *(Baseline)*	0.2 ± 0.1	11.2 ± 9.3	38.3 ± 12.0	$\mathbf{54.9} \pm 7.6$	26.2
FewNERD$_{\text{INTER}}$	LitSet	$\mathbf{24.3} \pm 0.6$	$\mathbf{39.8} \pm 2.9$	$\underline{49.1} \pm 1.9$	$\underline{52.1} \pm 1.9$	$\mathbf{41.3}$
	w/ all labels	$\underline{17.6} \pm 2.5$	36.1 ± 4.7	47.2 ± 3.0	50.4 ± 2.4	37.8
	w/ labels only	2.9 ± 0.6	24.7 ± 1.8	37.9 ± 1.7	42.4 ± 2.0	27.2
	w/ description only	16.2 ± 2.0	37.4 ± 2.9	47.8 ± 2.2	50.9 ± 1.9	38.1
	FewNERD$_{\text{INTER}}$ *(Baseline)*	10.6 ± 0.8	$\underline{38.4} \pm 3.1$	$\mathbf{50.4} \pm 3.1$	$\mathbf{53.3} \pm 2.6$	$\underline{38.2}$

Table 2: Evaluation of zero- and few-shot tagset extension for three datasets (FewNERD$_{\text{INTRA}}$, Ontonotes, FewNERD$_{\text{INTER}}$). We compare the baseline approach of using in-domain data for label interpretation learning against using LitSet. Despite lacking the in-domain advantage of the baselines, training on LitSet matches or significantly outperforms the in-domain baseline in nearly all settings. Best scores in bold, 2nd best underlined.

we note that due to our sampling method, LitSet annotations may not always be consistent. Thus, we can only ensure excluding exact overlaps with the few-shot domain.

4.1.2 Results

The experimental results are shown in Table 2 and find that LitSet substantially improves the few-shot performance in in-domain settings.

Detecting general entity types. We first observe that classifying completely new entity types is difficult with existing datasets (cf. OntoNotes and FewNERD (INTRA)). Even though masking all target labels and the limited exposure to in-domain data, our approach can effectively leverage its general label interpretation ability to strongly outperform baselines. We report +14.8 F1 on average in .pp on FewNERD$_{\text{INTRA}}$ and +3.3 F1 on OntoNotes. While LitSet consistently outperforms FewNERD (INTRA) except when k = 10 in the OntoNotes setting.

Differentiating coarse entity types. When coarse entity types are learned during label interpretation training (cf. FewNERD$_{\text{INTER}}$), we observe that all approaches obtain improved few-shot capabilities, especially when k < 5. This finding suggests that adapting to unseen labels is particularly effective when the training includes understanding broad categories (e.g., "person"). With LitSet, we outperform FewNERD$_{\text{INTER}}$ in 0- and 1-shot settings (+13.7 F1 and +1.4 F1 on average in pp.) and remain competitive at higher k-shots.

Impact of label semantics. We measure the impact of different heuristics for creating LitSet types. To test this, we conduct various experiments using LitSet with (*1*) only labels, (*2*) only descriptions, and (*3*) all label information available (cf. Figure 3). We first find using only label annotations results in decreased performance compared to the baselines (cf. FewNERD$_{\text{INTER}}$ and OntoNotes), suggesting the need for richer label meanings.

When using only the `description` annotations, we notice that LitSet yields similar performance to their respective baselines, whereas in the FewNERD$_{\text{INTRA}}$ setting, substantial improvements are observed compared to the baselines.

At last, we observe that alternating shorter labels and expressive short descriptions best prepares LitSet for arbitrary target domains. In this configuration, we find that LitSet substantially outperforms all baselines.

4.2 Experiment 2: Cross-Domain Transfer

This experiment assesses the performance of LitSet and its corresponding baselines when domains of label interpretation learning and few-shot fine-tuning differ. We selected out-of-domain datasets to cover labels that are not present in the current

Evaluation data $\mathcal{D}^{FS}$ for tagset extension from:	Label interpretation learning data $\mathcal{D}^{LIT}$ from:	0-shot	1-shot	5-shot	10-shot	Avg.
JNLPBA	LitSet	41.3 $\pm$ 2.0	25.4 $\pm$ 5.3	**51.3** $\pm$ 3.4	**57.7** $\pm$ 3.0	**43.9**
	w/ all labels	**42.2** $\pm$ 1.8	22.5 $\pm$ 8.1	49.9 $\pm$ 3.8	55.8 $\pm$ 2.7	42.6
	FewNERD$_{\text{INTER}}$	8.2 $\pm$ 1.5	**29.5** $\pm$ 15.0	46.0 $\pm$ 7.6	49.7 $\pm$ 6.6	33.4
CLUB	LitSet	6.1 $\pm$ 0.9	19.4 $\pm$ 3.3	25.9 $\pm$ 3.7	33.0 $\pm$ 2.1	21.1
	w/ all labels	**7.3** $\pm$ 0.1	**19.9** $\pm$ 2.0	**27.6** $\pm$ 4.6	**35.1** $\pm$ 3.1	**22.5**
	FewNERD$_{\text{INTER}}$	1.7 $\pm$ 0.2	16.9 $\pm$ 1.8	25.5 $\pm$ 4.9	32.2 $\pm$ 3.7	19.1

Table 3: LitSet outperforms FewNERD in out-of-domain settings on JNLPBA (bio-medical domain) and CLUB (chemical domain).

NER dataset to assess the genuine few-shot aspect of these models. We compare our approach with FewNERD$_{\text{INTER}}$ in this context. The results are presented in Table 3.

4.2.1 Experimental Setup

For out-of-domain experiments, we utilize JNLPBA (Collier et al., 2004) (bio-medical domain) and the Chemical Language Understanding Benchmark (CLUB) (Kim et al., 2023) (chemical domain). As detailed in Appendix C, our approach demonstrates transferability to datasets beyond those used in this experiment. However, we excluded them from our analysis here due to their limited number of distinct entity types and their label overlap with baseline models.

4.2.2 Results

As Table 3 shows, we find that LitSet significantly outperforms FewNERD with average improvements of +10.5 F1 on JNLPBA and +3.4 F1 on CLUB.

LitSet better transfers to new domains. While our approach consistently outperforms FewNERD on CLUB and JNLPBA for k-shot > 5, LitSet achieves an average increase of +34.0 F1 pp. in zero-shot settings on JNLPBA. This notable improvement can be attributed to the equal masking procedure applied to labels in FewNERD$_{\text{INTER}}$ and LitSet. Since JNLPBA labels and FewNERD labels are disjoint, no additional masking is required for FewNERD$_{\text{INTER}}$ models. Consequently, to maintain a fair comparison, we do not mask any labels in LitSet.

Impact of inconsistent annotations. Furthermore, we observed that LitSet underperforms by -4.1 F1 pp. compared to the baseline in 1-shot settings on JNLPBA. Additionally, its performance is inferior even compared to the 0-shot scenario. This indicates the instability of few-shot fine-tuning with LitSet at very low k. Upon further qualitative analysis of the generated dataset, we discovered that annotations from entity linking benchmarks like ZELDA might not be consistently annotated (cf. Appendix F). This inconsistency could be one possible reason for the observed performance drops. However, as k increases, our approach demonstrates the ability to quickly adapt to the target domain once again.

4.3 Experiment 3: Cross-Lingual Transfer

In this experiment, we utilized the multilingual `xlm-roberta-base` model to assess the transferability of LitSet across languages. English OntoNotes was employed as the baseline for label interpretation training since ZELDA is an English corpus. The results are shown in Table 4.

Results. We find strong improvements across all k-shots on the *Arabic* and *Chinese* segments of OntoNotes, namely +3.9 F1 and +9.0 F1 on average in pp., respectively. These findings underscore our model's ability to discern subtle annotation differences across languages despite the similar contexts between label interpretation learning and few-shot fine-tuning in the baseline. This emphasizes our model's robust understanding of labels in multilingual scenarios.

Furthermore, we observed that utilizing `xlm-roberta-base` also improves LitSet's performance in monolingual settings, as discussed in Section 4.1. We were able to reduce the previous performance gap at k = 10 from -6.5 F1 to -0.5 F1 on average in pp., thereby increasing the overall performance from +3.3 F1 to +6.5 F1.

5 Related Work

Despite advancements achieved through pre-trained word embeddings (Peters et al., 2018; Akbik et al., 2018; Devlin et al., 2019; Liu et al., 2019;

Evaluation data $\mathcal{D}^{FS}$ for tagset extension from:	Label interpretation learning data $\mathcal{D}^{LIT}$ from:	0-shot	1-shot	5-shot	10-shot	Avg.
OntoNotes (EN)	LITSET (EN)	**9.9** ± 3.2	**27.4** ± 8.5	**46.4** ± 6.7	55.5 ± 6.4	**34.8**
	OntoNotes (EN)	0.3 ± 0.1	15.9 ± 8.4	41.1 ± 15.0	**56.0** ± 12.7	28.3
Ontonotes (AR)	LITSET (EN)	0.0 ± 0.0	**7.2** ± 6.1	**14.8** ± 6.3	**22.0** ± 5.8	**14.7**
	Ontonotes (EN)	0.0 ± 0.0	4.7 ± 4.7	12.8 ± 4.8	14.9 ± 7.9	10.8
Ontonotes (ZH)	LITSET (EN)	**3.0** ± 0.9	**22.7** ± 8.6	**37.6** ± 5.0	**42.8** ± 5.0	**26.5**
	Ontonotes (EN)	1.6 ± 0.3	10.8 ± 5.9	26.2 ± 6.9	31.2 ± 7.9	17.5

Table 4: Tag set extension with baseline pre-finetuning and few-shot fine-tuning in the same domain. LITSET outperforms models that are pre-finetuning on in-domain data when pre-finetuning is done on a small number of labels.

Yamada et al., 2020; Raffel et al., 2020), few-shot NER focuses explicitly on generalizing to previously unseen label categories by leveraging a small number of labeled examples.

Metric learning (Vinyals et al., 2016; Snell et al., 2017) is a common approach for few-shot NER (Fritzler et al., 2019; Wiseman and Stratos, 2019; Ziyadi et al., 2020) and employs a distance metric to learn a shared representation space and assign labels based on class prototypes (Yang and Katiyar, 2020; Hou et al., 2020; Ma et al., 2022a; Han et al., 2023). Additional components like contrastive loss (Das et al., 2022; Layegh et al., 2023) or meta-learning (de Lichy et al., 2021; Ma et al., 2022c; Wang et al., 2022a) often further improve the performance. Our approach aligns with this research direction because we employ the bi-encoder architecture as proposed in Ma et al. (2022a); Zhang et al. (2023) with an adapted loss calculation. However, prior work did not investigateimpact of the dataset used for label interpretation learning. We instead increase the richness of the training signal learning label verbalizations. Our approach may thus be applied to all prior work that relies on label verbalizations, but may require architectural adaptations to accommodate arbitrary labels.

Template-filling and prompting methods with (large) language models(Lewis et al., 2020; Brown et al., 2020; Raffel et al., 2020; Scao et al., 2023; Touvron et al., 2023) have been widely used in few-shot NER (Cui et al., 2021; Ma et al., 2022b; Lee et al., 2022; Chen et al., 2022b; Kondragunta et al., 2023; Ma et al., 2023) tasks. However, these approaches, relying on masked language model (MLM) objectives, may not be directly comparable to our method due to the scale of our labels. In its basic form, the template-based approach requires one forward pass per label or is limited by the model's maximum sequence length. Additionally, our approach does not depend on large language models, which are often unavailable or impractical for few-shot NER tasks.

While specific efforts have been made to adapt to tags in few-shot domains (Hu et al., 2022; Ji et al., 2022), these studies evaluated only a limited number of labels. Our approach shares similarities with (Ren et al., 2022) and Chen et al. (2022a), where models were pre-trained using event mentions and entity links, respectively. However, our approach differs significantly. In Ren et al. (2022), the pre-training objective targets at latent typing of entities, whereas our approach focuses on explicitly scaling up entity typing of few-shot NER models. Our distinction from Chen et al. (2022a) lies in our exploration of the effectiveness of distantly supervised training in a genuine few-shot context, wherein classes are not observed during label interpretation training.

6 Conclusion

This paper introduces LITSET, a novel approach for label interpretation training with a large-scale set of entity types. We utilize an entity linking dataset annotated with WikiData information, resulting in a dataset with significantly more distinct labels. We then conducted a thorough heuristical, data-based optimization of few-shot NER models using this dataset. Our experiments demonstrate that LITSET consistently outperforms various in-domain, cross-domain, and cross-lingual baselines. For example, we surpass FewNERD by +14.7 F1 on average in pp. and Chinese OntoNotes by +9.0 F1 on average in pp. in low-resource settings. Our method and experiments provide valuable insights into the factors influencing the performance of few-shot NER models utilizing label semantics.

Limitations

Our heuristic data-based optimization was an initial exploration to understand the impact of scaling the number of distinct entity types during label interpretation learning on few-shot capability. Given our focus on this optimization, we selected a commonly used backbone architecture and one entity linking dataset. While substantial improvements were achieved, it's important to note that we did not explore all possible architectures and entity linking benchmarks. Thus, applying our approach with different model architectures and entity disambiguation datasets might yield significantly varied results. Further investigation is necessary to comprehensively understand how these factors interact and to develop more generalized few-shot NER models and comparable evaluation settings.

Additionally, achieving 0-shot capability on completely unseen tags remains challenging, especially in languages different from the one used for label interpretation training. This limitation highlights the need for future research and the exploration of innovative techniques to enhance the adaptability of few-shot NER models in 0-shot scenarios, enabling them to handle diverse domains and situations effectively.

Lastly, concerning LITSET, our best results were obtained by learning solely from in-batch instances. Although this strategy is commonly used in machine learning, there is substantial related work on learning from negatives, such as contrastive learning. We believe that exploring other architectures and loss functions, including those from contrastive learning, could potentially further improve our method.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jiawei Chen, Qing Liu, Hongyu Lin, Xianpei Han, and Le Sun. 2022a. Few-shot named entity recognition with self-describing networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5711–5722, Dublin, Ireland. Association for Computational Linguistics.

Yanru Chen, Yanan Zheng, and Zhilin Yang. 2022b. Prompt-based metric learning for few-shot ner.

Yanru Chen, Yanan Zheng, and Zhilin Yang. 2023. Prompt-based metric learning for few-shot NER.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.

Cyprien de Lichy, Hadrien Glaude, and William Campbell. 2021. Meta-learning for few-shot named entity recognition. In *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*, pages 44–58, Online. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of* the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

Elena V. Epure and Romain Hennequin. 2022. Probing pre-trained auto-regressive language models for named entity typing and recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1408–1417, Marseille, France. European Language Resources Association.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 993–1000, New York, NY, USA. Association for Computing Machinery.

Chengcheng Han, Renyu Zhu, Jun Kuang, FengJiao Chen, Xiang Li, Ming Gao, Xuezhi Cao, and Wei Wu. 2023. Meta-learning triplet network with adaptive margins for few-shot named entity recognition.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393, Online. Association for Computational Linguistics.

Jinpeng Hu, He Zhao, Dan Guo, Xiang Wan, and Tsung-Hui Chang. 2022. A label-aware autoregressive framework for cross-domain NER. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2222–2232, Seattle, United States. Association for Computational Linguistics.

Bin Ji, Shasha Li, Shaoduo Gan, Jie Yu, Jun Ma, Huijun Liu, and Jing Yang. 2022. Few-shot named entity recognition with entity-level prototypical network enhanced by dispersedly distributed prototypes. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1842–1854, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yunsoo Kim, Hyuk Ko, Jane Lee, Hyun Young Heo, Jinyoung Yang, Sungsoo Lee, and Kyu-hwang Lee. 2023. Chemical language understanding benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 404–411, Toronto, Canada. Association for Computational Linguistics.

Murali Kondragunta, Olatz Perez-de Viñaspre, and Maite Oronoz. 2023. Improving and simplifying template-based named entity recognition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 79–86, Dubrovnik, Croatia. Association for Computational Linguistics.

Amirhossein Layegh, Amir H. Payberah, Ahmet Soylu, Dumitru Roman, and Mihhail Matskin. 2023. Contrastner: Contrastive-based prompt tuning for few-shot ner.

Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2022. Good examples make a faster learner: Simple demonstration-based learning for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2687–2700, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. 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. 2019. Roberta: A robustly optimized bert pretraining approach.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022a. Label semantics for few shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1956–1971, Dublin, Ireland. Association for Computational Linguistics.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuanjing Huang. 2022b. Template-free prompt tuning for few-shot NER. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5721–5732, Seattle, United States. Association for Computational Linguistics.

Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. 2022c. Decomposed meta-learning for few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1584–1596, Dublin, Ireland. Association for Computational Linguistics.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples!

Marcel Milich and Alan Akbik. 2023. ZELDA: A comprehensive benchmark for supervised entity disambiguation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2061–2072, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Liliang Ren, Zixuan Zhang, Han Wang, Clare Voss, ChengXiang Zhai, and Heng Ji. 2022. Language model pre-training with sparse latent typing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1480–1494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and Matthias Gallé et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Peiyi Wang, Runxin Xu, Tianyu Liu, Qingyu Zhou, Yunbo Cao, Baobao Chang, and Zhifang Sui. 2022a. An enhanced span-based decomposition method for few-shot sequence labeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5012–5024, Seattle, United States. Association for Computational Linguistics.

Zihan Wang, Kewen Zhao, Zilong Wang, and Jingbo Shang. 2022b. Formulating few-shot fine-tuning towards language model pre-training: A pilot study on named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3186–3199, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sam Wiseman and Karl Stratos. 2019. Label-agnostic sequence labeling by copying nearest neighbors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5363–5369, Florence, Italy. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375, Online. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023. Optimizing bi-encoder for named entity recognition via contrastive learning. In *The Eleventh International Conference on Learning Representations*.

Morteza Ziyadi, Yuting Sun, Abhishek Goswami, Jade Huang, and Weizhu Chen. 2020. Example-based named entity recognition.

A FewNERD Label Semantics in Validation Experiment

An overview of the label semantics used in our validation experiment.

Original Label	Adapted Label
O	XO
location-GPE	PH
person-politician	EX
organization-education	CE

Table 5: Extract of random two letter labels for FewNERD.

Original Label	Adapted Label
O	XO
location-GPE	geographical social-political entity
person-politician	politician
organization-education	education

Table 6: Extract of short labels for FewNERD.

Original Label	Adapted Label
O	XO
location-GPE	geographical entity such as cities, states, countries, and political entities
person-politician	politicians such as presidents, senators, and other government officials
organization-education	education institutions such as schools, colleges, and universities

Table 7: Extract of long labels for FewNERD.

B WikiData labels

Given all entity mentions from the entity linking dataset, we source various information from WikiData in natural language and annotate those entities with it. In the following, we present the selected attributes along with their respective definitions, which will serve as our labels:

1. `x instance-of y`: Entity `x` is a particular example and instance of class `y`. For example, entity K2 is an instance of a mountain.

2. `y subclass-of z`: Instance `y` is a subclass (subset) of class `z`. For example, instance class volcano is a subclass of a mountain.

3. `description`: A short phrase designed to disambiguate items with the same or similar labels.

We note that the `instance-of` and `subclass-of` categories commonly encompass multiple tags rather than being limited to a single tag, as demonstrated in the example in Figure 3. We also refer to **??** for information on filtering improper information obtained by WikiData.

C Transfer on Additional Datasets

In this ablation, we show that our approach also transfers to the well-known datasets of CoNLL and WNUT. However, we excluded such datasets from our main experiments due to their limited amount of distinct labels (e.g., 4 labels for CoNLL, 6 labels for WNUT).

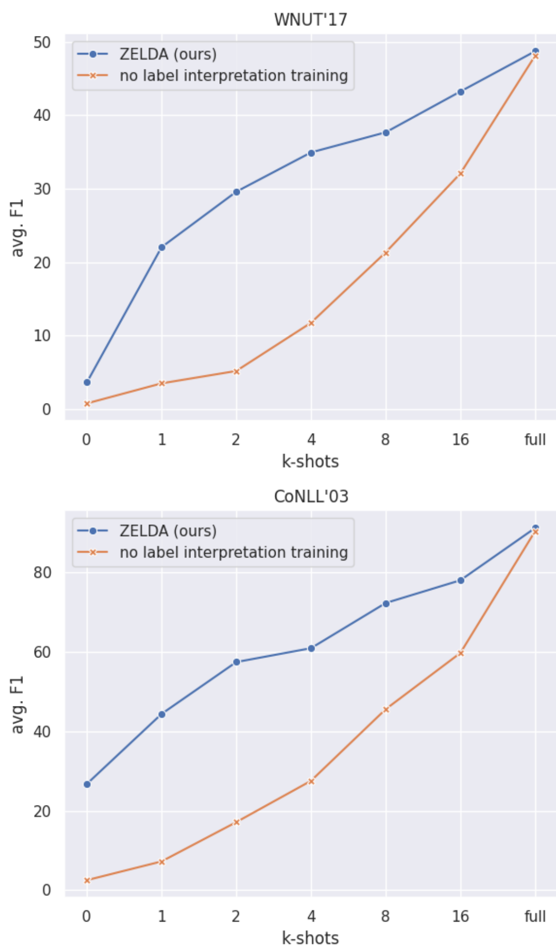

Figure 5: LITSET transfers to other datasets than the ones used in our main experiments. However, we excluded these datasets due to their limited number of distinct labels.

D Using Sentence-Transformers as Label Encoder

In this experiment, we investigate whether the sentence-transformer `all-mpnet-base-v2` can effectively help to better understand label semantics. Sentence transformers have been trained on a similarity objective, making them intriguing for our model to act as an enhanced label encoder. While LITSET performs consistently better compared to the baseline, we find that the standard sampling approach (using the `bert-base-uncased` transformer) works better.

E The Impact of Negative Examples

In this experiment, we investigate the impact of integrating negative labels $\mathcal{L}^-$ in each batch. To do so, we additionally sample negative labels from $\mathcal{L} \setminus \mathcal{L}_b$ until the desired number of labels is reached and include them for loss calculation, which could potentially lead to a better generalization in few-shot settings due to the increased signal during loss calculation. The results are shown in Table 9. We can observe that including more labels in each batch harms the performance. While prior work (Epure and Hennequin, 2022; Wang et al., 2022b) has shown that this is beneficial in few-shot settings, we find that LITSET works best when only using the label present in the batch for loss calculation. Since we randomly sample additional labels, it is possible, if not likely, to sample similar labels that are not true negatives and thus not advantageous when using cross-entropy loss.

F Annotation Noise in ZELDA

We find ZELDA, in some cases, is not consistently annotated which may effect the few-shot fine-tuning performance for in settings with very low k.

Evaluation data $\mathcal{D}^{FS}$ for tagset extension from:	Label interpretation learning data $\mathcal{D}^{LIT}$ from:	1-shot	5-shot	10-shot	Average
FewNERD$_{\text{INTRA}}$	FewNERD$_{\text{INTRA}}$	10.7 ± 7.4	37.8 ± 9.8	49.1 ± 8.4	32.5
	LITSET	$\mathbf{27.6} \pm 4.1$	$\mathbf{49.2} \pm 3.4$	$\mathbf{54.7} \pm 4.8$	**43.8**
FewNERD$_{\text{INTER}}$	FewNERD$_{\text{INTER}}$	23.4 ± 2.4	42.3 ± 3.8	$\mathbf{48.5} \pm 3.1$	38.1
	LITSET	$\mathbf{36.6} \pm 2.0$	$\mathbf{44.3} \pm 2.0$	47.7 ± 2.1	**42.9**

Table 8: Using sentence transformers as the label encoder. While ZELDA compares relatively better compared to the in-domain baseline, using sentence-transformers hurt the performance compared to the default `bert-base-uncased` transformer.

Evaluation data $\mathcal{D}^{FS}$ for tagset extension from: (/w # max. negative labels per batch)	Label interpretation learning data $\mathcal{D}^{LIT}$ from:	1-shot	5-shot	10-shot	Average
FewNERD$_{\text{INTRA}}$	LITSET (0)	$\mathbf{20.1} \pm 5.0$	$\mathbf{47.7} \pm 6.0$	$\mathbf{54.1} \pm 5.9$	**40.6**
	LITSET (64)	$\mathbf{20.1} \pm 4.8$	47.5 ± 5.0	53.2 ± 6.6	40.3
	LITSET (128)	18.9 ± 4.9	46.4 ± 3.9	52.7 ± 5.9	39.3
FewNERD$_{\text{INTER}}$	LITSET (0)	$\mathbf{36.1} \pm 4.7$	47.2 ± 3.0	50.4 ± 2.4	**44.6**
	LITSET (64)	35.2 ± 4.1	$\mathbf{47.4} \pm 2.6$	$\mathbf{50.5} \pm 2.4$	44.4
	LITSET (128)	34.7 ± 3.3	47.3 ± 2.7	50.4 ± 2.3	44.1

Table 9: The few-shot generalization of LITSET does not improve with a fixed number of labels per batch (we sample additional labels for loss calculation until, e.g., 64 labels are present). We find the best training setup to be only using the labels present in the current batch.

	Annotation noise in ZELDA
annotated	[...] which in turn creates the compound oxyhemoglobin | **protein** .
missing annotation	[...] whereas in oxyhemoglobin | **O** it is a high spin complex.
annotated	GSTK1 promotes adiponectin | **protein** multimerization
missing annotation	[...] ER stress induced adiponectin | **O** downregulation [...]

Table 10: Annotations in entity linking benchmark may be inconsistent, possibly causing the 1-shot drops on JNLPBA, given the dataset is human annotated, which should be consistent across all sentences.