

Learning from Language Description: Low-shot Named Entity Recognition via Decomposed Framework

Yaqing Wang[§], Haoda Chu[†], Chao Zhang[◇] and Jing Gao[§]

[§]Purdue University, [†]Microsoft, [◇]Georgia Institute of Technology
wang5075@purdue.edu, haochu@microsoft.com,
chaozhang@gatech.edu, jinggao@purdue.edu

Abstract

In this work, we study the problem of named entity recognition (NER) in a low resource scenario, focusing on few-shot and zero-shot settings. Built upon large-scale pre-trained language models, we propose a novel NER framework, namely SpanNER, which learns from natural language supervision and enables the identification of never-seen entity classes without using in-domain labeled data. We perform extensive experiments on 5 benchmark datasets and evaluate the proposed method in the few-shot learning, domain transfer and zero-shot learning settings. The experimental results show that the proposed method can bring 10%, 23% and 26% improvements in average over the best baselines in few-shot learning, domain transfer and zero-shot learning settings respectively.

1 Introduction

Named entity recognition (NER) aims at identifying and categorizing spans of text into a pre-defined set of classes, such as people, organizations, and locations. As a fundamental language understanding task, NER is widely adopted in question answering (Mollá et al., 2006), information retrieval (Guo et al., 2009) and other language understanding applications (Nadeau and Sekine, 2007; Ritter et al., 2012; Peng et al., 2020). Recent advances with pre-trained language models like BERT (Devlin et al., 2019a), GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019) have shown remarkable success in NER. However, the success of these large-scale models still relies on fine-tuning them on large amounts of in-domain labeled data. Unfortunately, obtaining NER annotations not only is expensive and time consuming, but also may not be feasible in many sensitive user applications due to data access and privacy constraints. This motivates us to study the problem of low-shot NER.

Low-shot NER focuses on identifying custom entities in a new domain with only a few in-domain

examples or even without any in-domain labeled data, which are referred to as few-shot NER and zero-shot NER respectively. The success of low-shot NER requires the model to be capable of transferring learned knowledge to recognize new entity classes. Conventional NER models usually treat each class as a one-hot vector (represented by a class label) for training, and thus the trained model cannot capture the semantic meanings of those labels. In fact, the trained model could be highly associated with known classes and it is difficult to transfer learned knowledge to novel entity classes.

To tackle this problem, several recent works (Huang et al., 2020; Yang and Katiyar, 2020; Hou et al., 2020; Ziyadi et al., 2020; Wiseman and Stratos, 2019) employ prototype-based method to represent each class by a prototype based on the labeled examples and use nearest neighbor method for NER. However, each entity class in NER task may include several fine-grained entity classes and has diverse semantic meanings. Correspondingly, the tokens or entities belonging to the same entity class are not necessarily close to each other (Huang et al., 2020), making it challenging to represent each entity class by a prototype based on a few examples. For example, MISC, one of the four entity classes in the benchmark dataset CoNLL03 (Sang and Meulder, 2003), is a collection of fine-grained entity classes including events, nationalities, products and works of art. FAC is an entity class in the OntoNotes5 (Weischedel et al., 2012) collection, including buildings, airports, highways, bridges and others. Thus, prototype-based methods may end up learning noisy representations of prototypes and cannot achieve satisfactory performance. Moreover, prototype-based methods have unavoidable reliance on labeled examples, thereby making them unable to extend to the zero-shot learning setting, which is also an important and practical scenario in the low-shot NER.

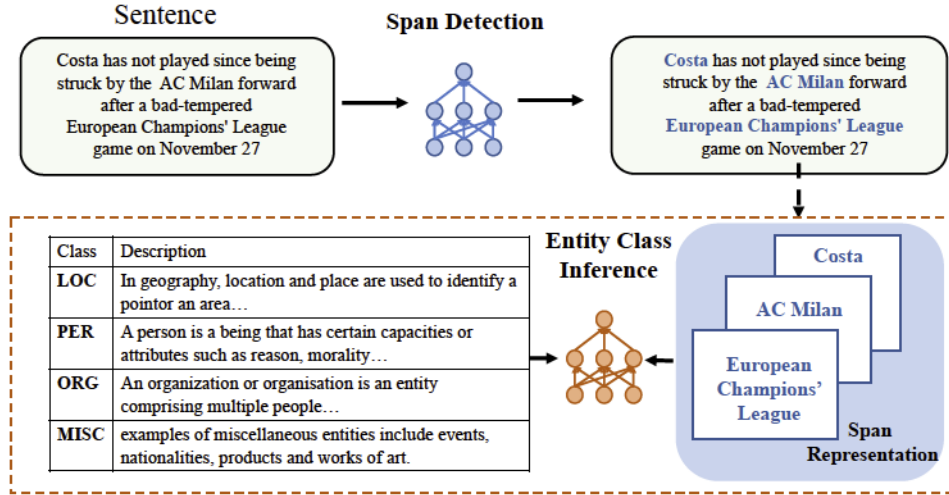


Figure 1: An overview of the proposed NER system: SpanNER.

In this paper, we propose a simple yet effective method SpanNER which can tackle few-shot as well as zero-shot NER. Instead of deriving the representations of labels purely from a few labeled examples, we propose to directly learn from the natural language descriptions of entity classes. Such a choice provides a flexible and precise way to obtain the semantic meanings of entity classes and enable zero-shot learning. Although using natural language as supervision has been explored in the context of zero-shot text classification, it is challenging to be adapted in the NER task. Unlike its use in text classification, natural language cannot provide direct supervision for token classification. Inspired by machine reading comprehension (MRC) framework, we propose to decompose the NER task into two procedures: span detection and entity class inference, which can be jointly trained together. However, it is challenging to employ MRC framework into NER task especially in the low-resource setting. The MRC framework usually needs a large amount of data for training, which is not available in the low resource scenario. To handle these challenges, we propose a class-agnostic span detection module which is equipped with token sampling mechanism to mitigate the imbalanced class issue and can be trained with limited labeled data. Moreover, to handle the challenging that several fine-grained classes are included in one entity class, we develop an entity class attention module to focus on the most relevant information in the given entity class description that corresponds to the extracted span. Compared with direct adaption of MRC framework into NER (Li et al., 2019), the proposed method can bring more

than 54%, 30% and 26% improvement in average under few-shot learning, domain transfer and zero-shot learning settings respectively. Figure 1 shows an overview of the proposed framework: the span detection stage aims to identify the span of text, and entity class inference is responsible for categorizing extracted spans based on natural language description of pre-defined entity classes. We perform extensive experiments on 5 benchmark datasets and evaluate the proposed method in the few-shot learning, domain transfer and zero-shot learning settings. The experimental results show that the proposed method brings large improvements over the state-of-the-art methods across different settings.

Contributions. Our model design is simple but distinguishes from that of the other NER works. To the best of our knowledge, we are the first one to learn entity class via natural language for NER task and the proposed method SpanNER achieves around 10%, 23% and 26% improvement over state-of-the-art NER methods in few-shot learning, domain-transfer and zero-shot learning settings respectively.

2 Task Formulation

NER is the process of locating and classifying named entities in text into predefined entity categories, such as person names, organizations, and locations. Formally, given a sentence with N tokens $X = \{x_1, \dots, x_N\}$, an entity or slot value draws from a span of tokens $s = [x_i, \dots, x_j] (0 \leq i \leq j \leq N)$ associated with an entity class $c \in \mathcal{C}$. The corresponding annotations for given sentence X are denoted as Y .

Few-shot NER focuses on the NER task in a low-

resource setting, where a system is only provided with a few in-domain labeled examples per entity class and the system needs to learn to identify custom entities in the domain. The task of K -shot NER refers to the setting with K labeled input sentences per entity class $c \in \mathcal{C}$, and the training data can be denoted as $\mathcal{D}_{\text{train}} = \{X_i, Y_i\}_{i=1}^{|\mathcal{C}| \times K}$. In this work, we leverage given training data $\mathcal{D}_{\text{train}}$ for model fine-tuning.

Zero-shot NER focuses on a more challenging setting where a model is trained with a set of entity classes and then tested on a dataset with a different set of entity classes. Towards this end, zero-shot NER systems need to learn to generalize to unseen entity classes without using any labeled example. The training data for zero-shot learning can be denoted as $\mathcal{D}_{\text{train}}$ associated with an entity class set $\mathcal{C}_{\text{train}}$, and the test data is denoted as $\mathcal{D}_{\text{test}}$ associated with an entity class set $\mathcal{C}_{\text{test}}$. Note that $\exists c_{\text{test}} \in \mathcal{C}_{\text{test}}$ but $c_{\text{test}} \notin \mathcal{C}_{\text{train}}$.

3 Methodology

In this paper, we study how to develop an effective model which can identify custom entities in a novel domain with a small set of labeled data or even without using any labeled data. To this end, we decompose NER task into two sub-tasks: span detection and entity class inference. The span detection module is class-agnostic and can transfer knowledge across different entity classes. On top of span detection, the entity class inference module takes extracted spans as input to infer the semantic relationship between the spans with natural language description of entity classes. Learning from natural language has an important advantage over existing categorical label learning methods, which is its ability to capture semantic meanings of labels and enable flexible zero-shot transfer.

3.1 Span Detection

Span detection is explored in the machine reading comprehension (MRC) frameworks (Chen et al., 2017; Seo et al., 2016), which predict the probability for each token as the starting or ending of the answer span given a question. However, it is challenging to directly adapt MRC framework for NER task especially in the low-resource setting. First, for each entity class, the model needs to answer its associated natural language question and repeat this procedure until all the questions are answered (Li et al., 2019). Thus, such a method is not scalable

when the number of entity classes increases and further exacerbates the imbalanced class issue compared to conventional NER framework. Second, the MRC framework usually needs a large amount of data, which is not available in the low resource scenario. To handle these challenges, we propose a class-agnostic span detection module which can share the knowledge across classes and develop a token sampling mechanism to mitigate imbalanced issue. The proposed span detection module takes input sequence as input without questions and can be trained with limited labeled data. Given an input sequence $X = \{x_1, \dots, x_N\}$, we first feed X into a pre-trained BERT (Devlin et al., 2019a) to obtain token representations $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \in \mathbb{R}^{h \times N}$. Besides start and end index predictions, we also classify whether a token is a part of the entity span. For example, we get the score for each token being start as follows:

$$s_{\text{start}}(i) = \mathbf{w}_{\text{start}}^T \cdot \mathbf{x}_i, \quad (1)$$

where $\mathbf{w}_{\text{start}} \in \mathbb{R}^{h \times 1}$ is the weight of the linear classifier. Correspondingly, the probability of a token being start index is:

$$p_{\text{start}}(i) = \text{sigmoid}(s_{\text{start}}(i)). \quad (2)$$

The probability calculations for end and a part of a span are the same as that of start index prediction. We then compute the probability of a span $[i, j]$ being an entity as:

$$p_{\text{match}}([i, j]) = \text{sigmoid}\left(s_{\text{start}}(i) + s_{\text{end}}(j) + \sum_{t=i}^j s_{\text{span}}(t)\right).$$

The span detection loss consists of three parts: start prediction loss, end prediction loss and span matching loss. The loss function of start prediction can be represented as:

$$\mathcal{L}_{\text{start}} = \frac{1}{N} \sum_{i=1}^N \text{CE}(p_{\text{start}}(i), y_{\text{start}}^i), \quad (3)$$

where CE represents cross-entropy function and $y_{\text{start}}^i = 1$ if token x_i is the start of an entity. The loss of end prediction can be calculated in a similar way.

Mitigation of imbalanced class issue. For an input sequence with length N , the number of span

candidate is in a $N \times N$ scale, where most of them are negative span candidates. To mitigate the imbalanced class issue, we sample a subset of negative span candidates, denoted as O^{neg} . The span candidate set which corresponds to gold spans is denoted as O^{pos} . Instead of using all the negative spans, we propose to sample a subset of negative spans to mitigate imbalance class issue. More specifically, we set the sampling size of negative span candidates as $|O^{\text{neg}}| = N - |O^{\text{pos}}|$ to reach a class ratio between positive and negative labels similar to the one in the start and end prediction losses. The role of such a mechanism is justified by an ablation study in Appendix. The span match loss is:

$$\mathcal{L}_{\text{match}} = -\frac{1}{N} \left(\sum_{(i,j) \in O^{\text{pos}}} \log p_{\text{match}}([i, j]) + \sum_{(i,j) \in O^{\text{neg}}} \log (1 - p_{\text{match}}([i, j])) \right).$$

The overall span objective consisting of three losses to be minimized is as follows:

$$\mathcal{L}_{\text{span}} = \mathcal{L}_{\text{start}} + \mathcal{L}_{\text{end}} + \mathcal{L}_{\text{match}}. \quad (4)$$

During inference, start and end indexes are first separately predicted. Then we select the consensus span between match predictions and extracted (start, end) indexes to achieve final predictions.

3.2 Natural Language Supervision

Learning based on categorical labels only may discard the semantic meanings of labels, and thus it is difficult to transfer knowledge from known classes to new entity classes. To mitigate this limitation, we propose to use natural language description¹ of entity classes to provide supervision for entity class inference and enable zero-shot learning. However, different with zero-shot text classification or entity linking, the entity class description in NER may describe several fine-grained entity classes, making our setup more challenging.

Mention Representation. Upon span detection, we can first obtain the mention span representation of each span candidate $[i, j]$ by averaging the embeddings of the span tokens. However, entity class is usually a high level category including many

entities. Thus, we need to add a linear transformation $\mathbf{w}_{\text{entity}} \in \mathbb{R}^{h \times h}$ to project average span token embedding into the entity class space:

$$\mathbf{e}_{i,j} = \mathbf{w}_{\text{entity}} \cdot \left(\frac{1}{j-i+1} \sum_{t=i}^j \mathbf{x}_t \right). \quad (5)$$

The description of an entity class c is a sequence of tokens, denoted as $X^c = \{x_1^c, \dots, x_K^c\}$. In this paper, we feed entity class description into another pre-trained BERT (Devlin et al., 2019b) to obtain its representations $\{\mathbf{x}_1^c, \dots, \mathbf{x}_K^c\} \in \mathbb{R}^{h \times K}$. Since there is limited data or even no data for training in the novel domain, we fixed the parameters of this BERT to expedite transferring by maintaining the embedding of entity class description from source and novel domains in the same space.

Entity Class Description Attention. The entity class description may describe several fine-grained entity classes. To focus on the information in the description that corresponds to the extracted span, we propose to construct adaptive entity class representation. More specifically, we use multi-headed attention mechanism (Vaswani et al., 2017). Each single attention function can be described as mapping a query and a set of key-value pairs to an output. The query, key and value vector are denoted as \mathbf{Q} , \mathbf{K} and \mathbf{V} respectively. We use the aggregated mention vector $\mathbf{e}_{i,j} \in \mathbb{R}^{h \times 1}$ as query vector \mathbf{Q} and use entity class description embedding $\mathbf{X}_c = [\mathbf{x}_1^c, \dots, \mathbf{x}_K^c] \in \mathbb{R}^{h \times K}$ as key vector \mathbf{K} and value vector \mathbf{V} . The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by the dot-product function of the query with the corresponding key. Then multiple parallel attention heads can stabilize the learning mechanism. We represent the procedure of obtaining adaptive entity class representation $\mathbf{x}^c(e_{i,j}) \in \mathbb{R}^{h \times 1}$ as:

$$\mathbf{x}^c(e_{i,j}) = \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}). \quad (6)$$

The role of such a mechanism is empirically justified by the comparison between the proposed model and a reduced model (i.e., the proposed model without attention mechanism) in the experimental section.

Entity Class Inference. The entity class inference is to infer the relationship between entity class and extracted span. We follow the zero-shot text classification (Yin et al., 2019) to cast this task into binary prediction: whether the extracted span belongs to given entity class or not. The probability

¹In this paper, we use the definitions of entity classes from Wikipedia or annotation guidelines as their language description.

of the extracted span being in a given entity class c is based on a matching score between them:

$$p(c|e_{i,j}) = \text{sigmoid}(\mathbf{e}_{i,j}^T \mathbf{x}^c(e_{i,j})). \quad (7)$$

The loss for each extracted span $[i, j]$ is calculated as:

$$\mathcal{L}_{\text{entity}}([i, j]) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{CE}(p(c|e_{i,j}), y), \quad (8)$$

where \mathcal{C} is a set of entity classes of interest, and y is binary label which equals to 1 when extracted span belongs to entity class c and 0 otherwise. We use $\mathcal{L}_{\text{entity}}$ to denote the entity class inference loss for all extracted spans.

Final Loss. We jointly train span detection and entity class inference modules by optimizing the sum of their losses.

4 Experiments

In this section, we empirically study and compare the proposed method with state-of-the-art methods in few-shot learning, domain transfer and zero-shot learning settings.

4.1 Experimental Setup

Dataset. We perform large-scale experiments with five different datasets² including Named Entity Recognition tasks and user utterances for task-oriented dialog systems as summarized in Table 1. (a) CoNLL03 (Sang and Meulder, 2003) is a collection of news wire articles from the Reuters Corpus with 4 entity classes. (b) OntoNotes5 (Weischedel et al., 2012) is in general domain including 18 entity classes. (c) WNUT 2017 (Derczynski et al., 2017) is collected from social media with 6 entity classes. (d) MIT Movie and Restaurant corpus (Liu et al., 2013) consist of user utterances for movie and restaurant domains with 12 and 8 classes.

Dataset	Domain	# Classes	# Train	# Test
CoNLL03	News	4	14K	3.6K
OntoNotes5	General	18	60K	8.3K
WNUT	Social Media	6	3.4K	1.6K
Movie	Moive	12	8.8K	2.4K
Restaurant	Restaurant	8	6.9K	1.5K

Table 1: Dataset summary.

Backbone We use the the pre-trained BERT_{base} uncased model ($\sim 110\text{M}$ parameters) as the backbone network. The inputs during training and inference

are lowercased to make them case-insensitive. The implementation details and hyper-parameter configurations are presented in Appendix.

4.2 Few-shot Learning

Setting. In this subsection, we study how the proposed method performs in a few-shot supervision setting. For 5-shot setting, we sample 5 sentences for each entity class from the training set and fine-tune models with sampled sentences. The experiment is repeated for 10 times to report the average F1 score. We also explore the role of distantly supervised learning in few-shot learning and the corresponding details are in Appendix.

Baselines The first baseline we use is a fully supervised BERT model trained on all available training data (3.4K-60K sentences) which provides the ceiling performance for every task. Each of the other models are trained on 5 training sentences per class. We compare our method with BERT (same backbone with ours) with Beginning-Intermediate-Outside (BIO) tagging mechanism as a comparison to evaluate the proposed model design besides the backbone choice. LC and Prototype are abbreviations for linear classifier and prototype-based methods from a recent few-shot NER work (Huang et al., 2020). They use pre-trained model RoBERTa-base as their backbone model. MRC-NER (Li et al., 2019) casts NER task into machine reading comprehension and achieves the state-of-the-art performance on several benchmark datasets. To study the role of attention mechanism proposed in subsection 3.2, we propose a reduced model SpanNER-NoAttn, which uses average operation instead of attention to aggregate entity class description.

Performance We report the results of 5-shot supervision and distantly supervising pre-training plus 5-shot supervision in Table 8. In the 5-shot supervision setting, we can observe that our methods outperform baseline BERT consistently, which shows the advantage of the proposed model design in addition to the benefits from backbone. The baseline Prototype leverages given support examples to conduct NER task and achieves lower performance compared with LC with the same backbone according to average F1. The reason may lie in that the tokens belonging to the same entity class are not necessarily close to each other (Huang et al., 2020). Prototype achieves better performance on WNUT compared to SpanNER since Prototype is based on Roberta which is pre-trained on social media

²<https://github.com/juand-r/entity-recognition-datasets>

Method	CoNLL03	OntoNotes5	WNUT	Movie	Restaurant	Average
Full-supervision						
BERT	91.1	87.8	47.1	87.9	79.0	78.6
5-shot supervision						
BERT	61.6	60.1	21.2	61.9	48.6	50.7
LC [†]	53.5	57.7	25.7	51.3	48.7	47.4
Prototype [†]	58.5	53.3	29.5	38.0	44.1	44.7
MRC-NER	28.5	49.8	0.4	58.7	43.1	36.1
SpanNER-NoAttn (ours)	<u>68.4</u> (0.5)	<u>65.1</u> (0.3)	22.8 (0.4)	<u>64.8</u> (0.3)	<u>48.9</u> (0.2)	<u>54.0</u>
SpanNER (ours)	71.1 (0.4)	67.3 (0.5)	<u>25.8</u> (0.3)*	65.4 (0.4)	49.1 (0.2)	55.7 [↑9.9%]

Table 2: F1 score comparison of models on different datasets. All models (except LC and Prototype) use the same BERT backbone. [†] indicates results from (Huang et al., 2020). The highest scores are **bolded**, while the second highest score is underlined. F1 score of our model for each task is followed by standard deviation and percentage improvement [↑] is over the best baseline. *Roberta is pre-trained on reddit dataset which is similar to WNUT. We change backbone from BERT to Roberta (same with Prototype’s) and F1 of SpanNER on WNUT is **31.5** (0.3).

dataset reddit. We change backbone of the proposed model SpanNER from BERT to Roberta-bas and observe that SpanNER achieves 31.5 in term of F1 score and outperforms Prototype. The MRC-NER framework is a reading comprehension framework whose success relies on training on large-scale data and thus cannot achieve satisfactory performance in a few-shot setting. Overall, we observe that our methods largely outperform all methods including the models with the same BERT encoder as ours across different datasets. The average performance improvement over the best baseline BERT is around 10%. Moreover, the comparison between SpanNER and SpanNER-NoAttn demonstrates that the improvement brought by attention mechanism is around 3.1%.

Varying the number of shots. Table 2 shows the improvement in the performance of SpanNER and BERT when increasing the number of labels for each NER type in the CoNLL03 dataset. As we increase the amount of labeled training instances, SpanNER improves over BERT consistently.

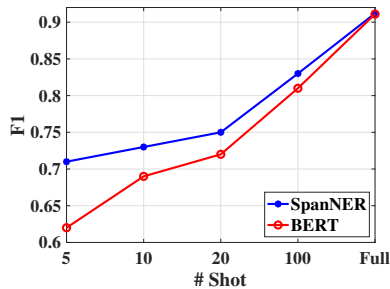


Figure 2: Variation in model performance on varying shots on CoNLL03. “Full” indicates full supervision.

4.3 Domain Transfer

We evaluate the proposed model in another common scenario of adapting a NER model to a novel domain (Yang and Katiyar, 2020). In this setting, we have a fully supervised source domain and a target domain with few-shot supervision. Following (Yang and Katiyar, 2020), we use general domain (OntoNotes5) as a source domain and evaluate models on News (CoNLL) and Social (WNUT) domains.

Models	CoNLL03	WNUT	Average
SimBERT [†]	28.6	7.7	18.2
Prototypical Network [†]	65.9	19.8	42.9
PrototypicalNet+P&D [†]	67.1	23.8	45.4
NNShot [†]	74.3	23.9	49.1
StructShot [†]	75.2	27.2	51.2
MRC-NER	64.1	32.6	48.3
SpanNER-NoAttn (ours)	80.1 (0.4)	42.0 (0.7)	61.1
SpanNER (ours)	83.1 (0.5)	43.1 (0.6)	63.1 [↑23.2%]

Table 3: F1 score comparison of models on CoNLL03 and WNUT datasets with 5-shot supervision for domain transfer. [†] indicates results from (Yang and Katiyar, 2020).

Baselines. We adopt six state-of-the-art methods in the domain transfer setting as baselines. SimBERT is based on a pre-trained BERT encoder and the predictions are conducted by a nearest neighbor classifier (Yang and Katiyar, 2020). Prototypical Network (Snell et al., 2017) is a state-of-the-art few-shot classification system and is adopted by (Fritzier et al., 2019) for few-shot NER task. Upon Prototypical Network, PrototypicalNet+P&D (Hou et al., 2020) adds pairwise embedding and dependency mechanism to gain further improvements. StructShot and NNShot are proposed in (Yang and Katiyar, 2020), which achieve state-of-the-art per-

formance in this domain transfer setting. We include our reduced baseline SpanNER-NoAttn in this task as an ablation study and follow (Yang and Katiyar, 2020) to run our models five times and report average performance with standard deviation. **Performance** Table 3 shows the results of baselines and the proposed methods on CoNLL03 and WNUT datasets. The proposed models achieve F1 scores 83.1 and 43.1 on CoNLL03 and WNUT respectively, bringing improvements over the best baseline 10.5% and 32.2% correspondingly. The proposed model effectively transfers learned knowledge to novel domains by learning from natural language descriptions instead of simple one-hot representation of entity classes.

4.4 Zero-shot NER

Setting The zero-shot learning setting is motivated by the fact that new types of entities often emerge in some domains and sometimes the annotations in the target domain are not accessible. Following zero-shot text classification setting (Yin et al., 2019), we evaluate the proposed model in a common setting: *label-partially-unseen*. In label-partially-unseen setting, a part of labels are unseen, enabling us to check the performance on unseen labels as well as seen labels.

Baselines Zero-shot NER task is rarely studied. The most state-of-the-art model for zero-shot NER is MRC-NER (Li et al., 2019), which conducts NER task by extracting answer spans given the questions of entity classes. Another baseline we use is the reduced model SpanNER-NoAttn. The comparison with this reduced model can demonstrate the role of entity class attention mechanism.

Method	CoNLL03		WNUT	
	Overall	Unseen	Overall	Unseen
MRC-NER	39.1	14.5	24.0	7.4
SpanNER-NoAttn	39.0	0.5	31.4	16.8
SpanNER	53.0	33.5	35.4	18.8

Table 4: F1 score comparison of models on CoNLL03 and WNUT datasets. *Overall* and *Unseen* indicate F1 scores of all entity classes and never-seen entity classes, respectively.

Performance Table 4 shows the F1 scores of MRC-NER and the proposed methods. MRC-NER based on reading comprehension framework is capable of conducting NER task for never-seen classes. However, the span detection in MRC-NER is tightly coupled with question understanding, leading to more difficulty in handling unseen entity classes. In

contrast, the proposed framework decomposes the NER task into span detection and entity class inference, avoiding the error propagation between two modules and thus delivering better performance. The comparison between SpanNER-NoAttn and SpanNER indicates the importance of attention mechanism in entity class description understanding, especially for never-seen entity classes.

Entity Class Inference We conduct experiments that disentangle entity class inference module from SpanNER so that the capability of this module can be evaluated. The span detection module cannot be separately evaluated because the span annotations on both datasets are associated with pre-defined entity classes. To demonstrate the capability of entity class inference, we use gold spans to evaluate the performance of entity class inference. Table 5 shows the performance of SpanNER-NoAttn and SpanNER. Comparing these two methods, we can observe that the attention mechanism helps improve the performance of the entity class inference.

	CoNLL03	WNUT
Entity Class Inference		
SpanNER-NoAttn	56.2	53.7
SpanNER	60.2	57.0
Annotation guidelines		
SpanNER-NoAttn	31.8	11.5
SpanNER	42.1	15.7

Table 5: Experiments that demonstrate the performance of the entity class inference module and adopt annotating guidelines as entity class descriptions on CoNLL03 and WNUT datasets.

Class Description Construction We set up experiments to study how the entity class description affects the model performance. In this experiment, we replace Wikipedia description of entity classes by annotation guidelines from CoNLL03 and WNUT datasets in the testing stage. We can observe that the proposed models are still capable of identifying entities belonging to never-seen entity classes even though the descriptions in the testing stage are different from those in the training stage. The F1 scores drop compared to the scores when Wikipedia descriptions are used because training and test stages use different descriptions and the annotation guidelines do not include the semantic explanation of entity classes.

Performance per Entity Class We show F1 score per entity class on CoNLL03 and WNUT datasets

in Table 6. We can observe the various degrees of recognizing different entity classes. First, the person names are easily recognized across different domains. The performance of person entity class on WNUT is worse compared to that on CoNLL03, which may be due to the large domain shift in social media data. It is interesting to see that the performance of seen entity classes `LOC`, `location` and `product` is even worse than that of never-seen entity classes. To explain this interesting phenomenon, we provide a detailed analysis of error cases below.

CoNLL03		WNUT	
Entity Class	F1	Entity Class	F1
PER	77.4	person	59.1
ORG	58.5	creative-work*	19.3
MISC*	33.5	corporation*	19.1
LOC	5.9	group*	18.3
-	-	location	14.4
-	-	product	11.0

Table 6: F1 score of SpanNER per entity class on CoNLL03 and WNUT datasets. * indicates unseen entity classes.

Error Analysis We manually examine the errors made by the proposed model on the CoNLL03 and WNUT test datasets and categorize these errors into 3 types. The error examples are presented in Appendix. (1) Different annotation guidelines on datasets. For instance, the description of `location` entity class in the source domain (OntoNotes5) is limited to mountain ranges and bodies of water, excluding countries, cities, states (these are included in the entity class `GPE`). Such a description is different from entity class `LOC` on CoNLL03 and `location` on WNUT. (2) Domain shift. The domain shift leads to the difficulty in recognizing the entities belonging to seen entity classes. (3) Description understanding. Description understanding is a crucial step for the success of zero-shot NER. For example, `MISC` on CoNLL03 is a collection of diverse fine-grained entity classes including events, nationalities, products and works of art.

5 Related Work

The most related topics are few-shot and zero-shot NER, which are discussed as follows.

Few-shot NER aims to build a model that can recognize a new class with a small number of labeled examples quickly. Recent works (Huang

et al., 2020; Yang and Katiyar, 2020; Hou et al., 2020; Ziyadi et al., 2020; Wiseman and Stratos, 2019) exploit prototype-based methods to conduct NER tasks. Since tokens or entities belonging to the same entity class are not necessarily close to each other, prototype-based methods usually end up learning noisy prototypes and may not achieve satisfactory performance. To further improve few-shot performance, (Hofer et al., 2018; Huang et al., 2020) explores different pre-training strategies for few-shot NER, and (Wang et al., 2020; Huang et al., 2020) propose to leverage self-training to take advantage of additional unlabelled in-domain data. Although aforementioned few-shot NER works show the potential of additional data in improving performance of few-shot NER, they still suffer from the limitations of prototype or one-hot representations of labels in transferring knowledge. Moreover, the aforementioned models cannot be applied in the zero-shot learning setting due to either reliance on labeled support set or the adoption of one-hot label representation.

Zero-shot NER is to build a model that can recognize new classes without using corresponding labeled data. This setting is rarely studied in NER task. Zero-shot NER is important and practical in the real scenario since the annotations may not be accessible due to privacy and compliance restrictions for some sensitive user applications. (Rei and Søgaard, 2018) has worked on zero-shot sequence labeling task by using attention to infer binary token-level labels. However, their token level predictions are constrained to being binary and has to rely on sentence labels. These limitations prohibit the use of this method for NER task. MRC-NER (Li et al., 2019) formulates NER task as a machine reading comprehension task and enables zero-shot NER. However, the inference of MRC-NER for the single sentence needs to be conducted multiple times to collect results corresponding to all the entity types of interest, incurring expensive inference cost. Moreover, the reading comprehension framework needs to be trained with large-scale dataset and is not effective in the few-shot setting. (Logeswaran et al., 2019; Wu et al., 2019) propose to incorporate entity description for zero-shot entity-linking task. Nevertheless, they are different with ours in several aspects. First, the input to this work are sentences with pre-annotated entities while our task takes sentences as input and needs to conduct entity recognition and entity type inference

jointly. Second, in the entity linking problem, the mentions in the sentences and Wikipedia descriptions refer to the same entities, while mentions in the sentences are only instances of entity classes in the NER task. Toward those challenges, SpanNER introduces how to capture class description which usually includes many fine-grained classes for a further improvement. Other zero-shot problems studied in NLP involve text classification (Yin et al., 2019), entity typing (Zhou et al., 2018), word sense disambiguation (Kumar et al., 2019) and relation extraction (Levy et al., 2017). These problems have different settings and challenges compared to zero-shot NER.

6 Conclusions

In this work, we study how to conduct NER in a low resource setting (when there are few or zero labeled data). To this end, we develop a novel NER framework SpanNER that decomposes NER task into span detection and entity class inference. This framework enables zero-shot NER and improves the performance of few-shot NER by capturing semantic meanings of entity classes. Extensive experiments on 5 benchmark datasets and various settings demonstrate the effectiveness of the proposed model, particularly in the low-resource settings.

Acknowledgment

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work is supported in part by the US National Science Foundation under grant NSF-IIS 1747614. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.
- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Alexander Fritzler, Varvara Logacheva, and Maksim Kreto. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 993–1000.
- Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274.
- Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. 2018. Few-shot learning for named entity recognition in medical text. *arXiv preprint arXiv:1811.05468*.
- 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.
- Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2020. Few-shot named entity recognition: A comprehensive study. *arXiv preprint arXiv:2012.14978*.
- Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.
- Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*.
- Belinda Z Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. In *Proceedings of the 2020 Conference on Empirical Methods*

- in *Natural Language Processing (EMNLP)*, pages 6433–6441.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*.
- Xiao Ling and Daniel Weld. 2012. Fine-grained entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26.
- J. Liu, Panupong Pasupat, D. Cyphers, and James R. Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390.
- 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](#). *CoRR*, abs/1907.11692.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460.
- Diego Mollá, Menno Van Zaanen, Daniel Smith, et al. 2006. Named entity recognition for question answering.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguistic Investigations*, 30(1):3–26.
- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayan-deh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pre-trained auto-regressive model. *arXiv preprint arXiv:2005.05298*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Marek Rei and Anders Søgaard. 2018. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. 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 293–302.
- Alan Ritter, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112.
- 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*.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.
- Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4080–4090.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2020. Adaptive self-training for few-shot neural sequence labeling. *arXiv preprint arXiv:2010.03680*.
- Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D Hwang, Claire Bonial, et al. 2012. Ontonotes release 5.0.
- 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.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.
- Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. *arXiv preprint arXiv:2010.02405*.
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3905–3914.
- Ben Zhou, Daniel Khashabi, Chen-Tse Tsai, and Dan Roth. 2018. Zero-shot open entity typing as type-compatible grounding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2065–2076.
- M. Ziyadi, Yuting Sun, A. Goswami, Jade Huang, and W. Chen. 2020. Example-based named entity recognition. *ArXiv*, abs/2008.10570.

A Implementation

Training We use AdamW (Devlin et al., 2019b) as the optimizer. The learning rate is selected from $\{6e-6, 1e-5, 2e-5\}$ and batch size is selected from $\{8, 16\}$ according to validation set, coupled with a linear schedule with 1% warmup. We clip gradients to max norm 1.0. For all training data sizes, we set the training epoch as 50. The max sequence length for input is 128 and the max sequence length for label description is 32. The dimension h is 768, which is the embedding dimension of BERT. The pre-trained language model for label description is a BERT_{base} uncased model. For multi-head attention mechanism, the multi-head number is 4, the hidden dimension size is 300 and the Dropout probability is 0.2. The model is run on 4 NVIDIA Titan Xp GPU servers.

Inference For few-shot NER, the inference of span detection is introduced in Subsection 3.1. Since the entity class inference is binary prediction for each entity class, we use 0.5 as decision boundary to determine whether the extract span belongs to given entity class or not. During the zero-shot NER, the mentions corresponding to never-seen classes may have lower detection scores compared to seen entity classes. Thus, we calculate the entity inference score using softmax over all the entity classes to get a “relative” score. Since many extracted spans may not be associated with the entity classes of interest, we follow (Li et al., 2020) to use a hyper-parameter threshold γ to select identified entities based on the joint score of span detection and entity class inference for span $[i, j]$. More specially, the joint score of span $[i, j]$ can be calculated by $\log(p_{\text{match}}([i, j])) + \log p(c|e_{i,j})$. The thresholds of zero-shot are selected as -0.4 and -0.5 for CoNLL03 and WNUT respectively according to validation set. We investigate the effect of γ in the Figure 3.

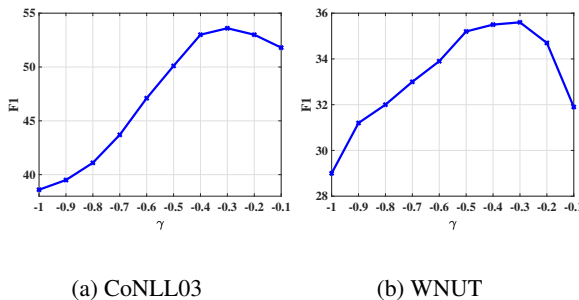


Figure 3: F1 score changes of our model w.r.t. varying threshold γ values.

Threshold γ We show the F1 score changes of the proposed method on test datasets of CoNLL03 and WNUT with respect to varying γ values in Figure 3. Similar trends can be observed on both datasets. The F1 scores first increase as threshold γ values increase and then reach the peak in the middle points.

B Token Sampling

Token sampling is introduced into the proposed model to mitigate the imbalance class issue, which degrade the efficiency and effectiveness of training especially in the low-resource scenario. To explore the role of token sampling, we conduct ablation study by removing the token sampling. The proposed baseline is denoted as SpanNER-NoSampling. We show the results on CoNLL03 and OntoNotes5 with 5-shot supervision in Table 7. Table 7 shows that the token sampling mechanism is effective and brings improvements around 4.7% and 2.3% on CoNLL03 and OntoNotes5 respectively.

	CoNLL03	OntoNotes5
SpanNER-NoSampling	67.9 (0.2)	65.8 (0.3)
SpanNER	71.1 (0.4)	67.3 (0.5)

Table 7: F1 score comparison of models on CoNLL03 and OntoNotes5 with 5-shot supervision. Overall highest scores are **bolded**. F1 score of models is followed by standard deviation.

C Distant Supervision

We are also interested in the role of distantly supervised pre-training in the few-shot NER task, which is studied in a recent work (Huang et al., 2020). We construct a Wikipedia distantly supervised dataset which contains 20 entity classes and around 1 million sentences. In this setting, we first pre-train NER models with distantly supervised data and then fine-tune the pre-trained models with 5-shot supervision.

To construct the training data, we use the May 2019 English Wikipedia dump and use the anchor text as the mention. We select 20 popular entity types from the fine-grained entity typing dataset Figer (Ling and Weld, 2012) including Locations, Organization, Person, City, Artist, Country, Author, Actor, Company, Event, Government, Sports team, Athlete, Title, Cemetery, Musician, Province, Building, Language, Politician. We retrieved entities belonging to those entity types by querying

Method	CoNLL03	OntoNotes5	WNUT	Movie	Restaurant	Average
Full-supervision						
BERT	91.1	87.8	47.1	87.9	79.0	78.6
Distantly Supervised Pretraining + 5-shot supervision						
BERT	65.9	64.9	36.2	64.1	50.7	56.4
LC [†]	61.4	68.8	34.2	53.1	49.1	53.3
Prototype [†]	60.9	57.0	35.9	43.8	48.4	49.2
MRC-NER	69.4	63.4	36.3	64.6	48.7	56.5
SpanNER-NoAttn (ours)	74.4 (0.3)	66.5 (0.3)	35.8 (0.6)	65.2 (0.2)	49.3 (0.2)	58.3
SpanNER (ours)	75.6 (0.4)	71.6 (0.2)	38.5 (0.4)	67.8 (0.3)	51.2 (0.1)	60.9 [\uparrow 7.8%]

Table 8: F1 score comparison of models on different datasets. All models (except LC and Prototype) use the same BERT backbone. [†] indicates results from (Huang et al., 2020). The overall highest scores are **bolded**. F1 score of our model for each task is followed by standard deviation. Percentage improvement [\uparrow] is over the best baseline in the corresponding setting.

from Wikidata using SPARQL. Then we use string matching to assign entity types to anchor text. We use a subset of the wikipedia data around 1M examples for training and 10K hold-out examples for validation.

The distantly supervised pre-training generally improves the performance for all the methods, confirming its potential in the low-resource NER task. Our model achieves 60.9 in average F1 and its average performance is improved by around 9.3% compared to that in the setting of 5-shot supervision. Overall, the proposed model achieves the best performance on all the datasets and the improvement of SpanNER over the best baseline is around 7.6% in average terms of F1.

D Error Cases

we show error examples made by our model in Table 9. In the first example on CoNLL03, “Japan” is annotated as LOC. However, the entity class location on OntoNotes5 (source domain) does not include countries and thus SpanNER does not classify this span into LOC. We also observe some incorrect predictions made by our model look reasonable. For instance, “Outagamie County Circuit Court” is identified as an entity belonging to ORG. Moreover, we find that our model correctly identifies “Luebke” as a PER entity which is not annotated in the dataset. On WNUT dataset, we find that many entities are difficult to identify especially that its context is very different with that of our source domain (OntoNotes5). For instance, “Watch What Else is Making News” only includes limited context words “Watch”.

Dataset	Example Error
CoNLL03	<p>Japan then laid siege to the Syrian penalty area for most of the game but rarely breached the Syrian defence.</p> <p>Gold: Japan → LOC</p> <p>Gold: Syrian → MISC</p> <p>Gold: Syrian → MISC</p>
CoNLL03	<p>Japan then laid siege to the Syrian penalty area for most of the game but rarely breached the Syrian defence.</p> <p>Pred: Syrian → MISC</p> <p>Pred: Syrian → MISC</p>
CoNLL03	<p>In sentencing Darrel Voeks, 38, to a 10-year prison term on Thursday, Outagmie County Circuit Court Judge Dennis Luebke said he was "a thief by habit."</p> <p>Gold: Darrel Voeks → PER</p> <p>Gold: Outagmie County → LOC</p> <p>Gold: Dennis Luebke → PER</p>
CoNLL03	<p>In sentencing Darrel Voeks, 38, to a 10-year prison term on Thursday, Outagmie County Circuit Court Judge Dennis Luebke said he was "a thief by habit."</p> <p>Pred: Darrel Voeks → PER</p> <p>Pred: 38 → MISC</p> <p>Pred: Outagmie County Circuit Court → ORG</p> <p>Pred: Dennis Luebke → PER</p>
CoNLL03	<p>You are narcissitic, Luebke said at the sentencing, adding Voeks should pay restitution of more than \$100,000 to the farming family who had hired him.</p> <p>Gold: Voeks → PER</p>
CoNLL03	<p>You are narcissitic, Luebke said at the sentencing, adding Voeks should pay restitution of more than \$100,000 to the farming family who had hired him.</p> <p>Pred: Luebke → PER</p> <p>Pred: Voeks → PER</p>
CoNLL03	<p>This cannot endure, "Marlow told BBC television's Newsnight programme on Thursday."</p> <p>Gold: Marlow → PER</p> <p>Gold: BBC → ORG</p> <p>Gold: Newsnight → MISC</p>
CoNLL03	<p>This cannot endure, "Marlow told BBC television's Newsnight programme on Thursday."</p> <p>Pred: Marlow → PER</p> <p>Pred: BBC television's → ORG</p>
WNUT	<p>Visuals of the avalanche site in Gurez sector.</p> <p>Gold: Gurez sector → location</p>
WNUT	<p>Visuals of the avalanche site in Gurez sector.</p> <p>Pred: Gurez → location</p>
WNUT	<p>Watch What Else is Making News</p> <p>Gold: What Else is Making News → creative-work</p>
WNUT	<p>Watch What Else is Making News</p>
WNUT	<p>[Rip Chad] (https://www.reddit.com/r/soccer/comments/5mi9bl/granada_and_...)</p> <p>Gold: Chad → person</p>
WNUT	<p>[Rip Chad] (https://www.reddit.com/r/soccer/comments/5mi9bl/granada_and_...)</p> <p>Pred: Rip Chad → person</p>
WNUT	<p>Step 2: [Google] (http://www.google.co.nz)</p> <p>Gold: Google → creative-work</p>
WNUT	<p>Step 2: [Google] (http://www.google.co.nz)</p> <p>Gold: Google → corporation</p>

Table 9: Error examples made by our model SpanNER in zero-shot NER setting. ■ denotes gold spans and ■ denotes predicted spans.