Few-shot Named Entity Recognition via Superposition Concept Discrimination

Jiawei Chen[1,3], **Hongyu Lin**[1,†], **Xianpei Han**[1,2],
Yaojie Lu[1], **Shanshan Jiang**[4], **Bin Dong**[4], **Le Sun**[1,2]
[1]Chinese Information Processing Laboratory, [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Ricoh Software Research Center Beijing Co., Ltd
{jiawei2020, hongyu, xianpei, yaojie, sunle}@iscas.ac.cn,
{shanshan.jiang, bin.dong}@cn.ricoh.com

Abstract

Few-shot NER aims to identify entities of target types with only limited number of illustrative instances. Unfortunately, few-shot NER is severely challenged by the intrinsic precise generalization problem, i.e., it is hard to accurately determine the desired target type due to the ambiguity stemming from information deficiency. In this paper, we propose Superposition Concept Discriminator (SuperCD), which resolves the above challenge via an active learning paradigm. Specifically, a concept extractor is first introduced to identify superposition concepts from illustrative instances, with each concept corresponding to a possible generalization boundary. Then a superposition instance retriever is applied to retrieve corresponding instances of these superposition concepts from large-scale text corpus. Finally, annotators are asked to annotate the retrieved instances and these annotated instances together with original illustrative instances are used to learn FS-NER models. To this end, we learn a universal concept extractor and superposition instance retriever using a large-scale openly available knowledge bases. Experiments show that SuperCD can effectively identify superposition concepts from illustrative instances, retrieve superposition instances from large-scale corpus, and significantly improve the few-shot NER performance with minimal additional efforts.

Keywords: few-shot learning, named entity recognition, active learning

1. Introduction

Few-shot named entity recognition (FS-NER) aims to detect and classify named entity from text with only a few illustrative instances. FS-NER is appealing for open-domain NER which contains various unforeseen types and very limited examples, and therefore has attached great attention in recent years (Fritzler et al., 2019; Yang and Katiyar, 2020; Ding et al., 2021; Huang et al., 2021).

Even with rapid progress, FS-NER faces severe intrinsic ***precise generalization*** challenge which is ignored by previous literature. Given only a few illustrative instances, it is frequently impossible to accurately determine what the desirable target entity type is. As a result, the learned FS-NER models often suffer from over-generalization or under-generalization. Figure 1 shows an illustrative example. Given two illustrative instances "Fei-Fei Li works in [Stanford University]." and "[Peking University] is located in Beijing.", we are unable to determine the target entity type is *University*, *EduIns*(*Educational institution*), *ResIns*(*Research institution*) or *Organization*. Consequently, if the target entity type is *EduIns*, a learned FS-NER model may unpredictably over-generalize to *Organization* or *ResIns*, it may also under-generalize to *University*. Note that precise generalization is a task-intrinsic challenge of FS-NER because it can not be addressed by designing better model architecture without introducing additional accurate generalization knowledge.

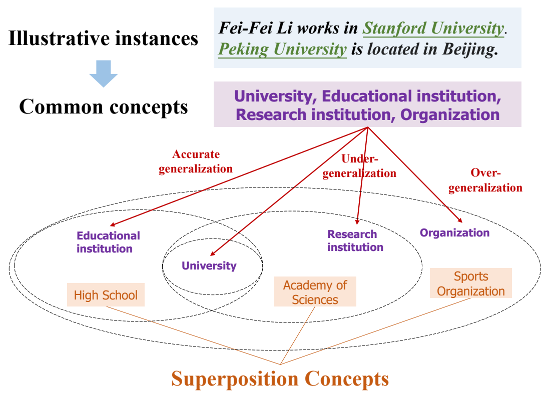


Figure 1: Examples of the precise generalization challenge. The underline indicates the annotated entity mentions. Given only the illustrative instances, the desired target type may be *University*, *Educational institution*, *Research institution* or *Organization*. Discriminating superposition concepts like *High school*, *Academy of sciences* and *Sports organization* helps determine what the desirable target entity type is.

The key to resolving the precise generalization

† Corresponding authors.

LREC-COLING 2024, pages 7220–7231

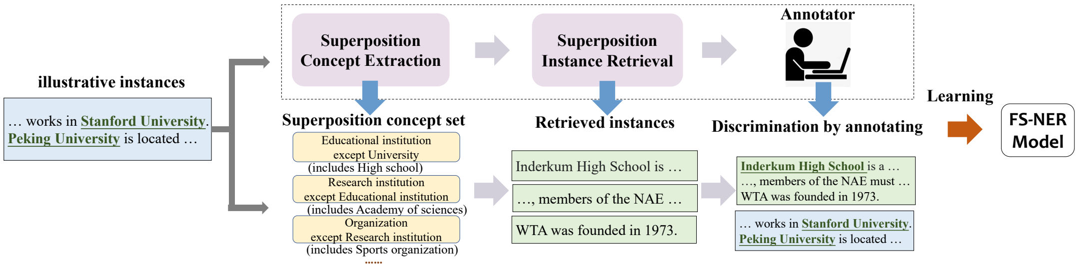


Figure 2: Overview of SuperCD. The underline indicates the annotated entity mentions. SuperCD first extract sets of superposition concepts and then retrieve corresponding instances. Finally, by annotating that Inderkum High School is the target type while NAE (*Research institution*) and WTA (Sports organization) is not and using these instances to learn model, the generalization knowledge is injected into the FS-NER model.

challenge is to provide information about critical ***superposition concepts***, which refers to the concepts that are associated with some common concepts entailed in illustrative instances. Without explicit declaration, instances of superposition concepts can both be or not be regarded as congener of illustrative instances. As a result, the target entity type can not be clarified unless providing sufficient measurement on these superposition concepts. For the example in Figure 1, *HSchool* (*High school*), *AcademySci* (*Academy of sciences*) and *SportsOrg* (*Sports organization*) are superposition concepts. On one hand, if we know that *HSchool* is part of the desirable type, we can avoid the under-generalization because we know that *University* is not the accurate target type. On the other hand, if we know that *AcademySci* and *SportsOrg* are not a desirable concepts, we can avoid over-generalizing to *ResIns* and *Organization*. As a result, a precise *EduIns* recognizer can be learned only by providing additional information about *HSchool*, *AcademySci* and *SportsOrg*. Unfortunately, accurately providing information about superposition concepts is very challenging due to the limited annotation budget and the unpredictable scale of potential superposition concepts. Consequently, simply annotating more instances is not only expensive but can not guarantee that the additionally-annotated instances can cover the superposition concepts that need to be measured. Therefore, how to identify critical superposition concepts and providing the information to FS-NER models with minimal additional efforts poses a huge challenge to FS-NER.[1]

In this paper, we propose Superposition Concept Discriminator (SuperCD), which resolves the above-mentioned challenge in an active learning (Lewis and Catlett, 1994; Settles, 2009; Shen et al., 2018; Zhou et al., 2021) paradigm. Figure 2 shows the overall framework of SuperCD. The main idea behind SuperCD is to introduce an elimination-based approach to identify sets of superposition concepts and leverage an instance retriever to find instances corresponding to superposition concepts from a large-scale text corpus. Annotators are then asked to annotate the retrieved instances and these annotated instances and original illustrative instances are used to learn FS-NER models. In this way, the high-value precise generalization knowledge entailed in the additional annotated instances can be injected into the FS-NER models. Specifically, to accurately identify the superposition concepts, we introduce a Concept Extractor (CE), and to retrieve instances of superposition concepts that need to be annotated, we introduce a Superposition Instance Retriever (SIR). Specifically, given a few illustrative instances of the target entity type, CE is first applied to extract the common concepts entailed in the illustrative instances. Then sets of superposition concepts are obtained based on an "A but not B" manner, in which the concept is part of concept A but not concept B. For the example in Figure 1, common concepts *University*, *Educational institution*, *Research institution* and *Organization* are first extracted, and then superposition concepts like *"University but not Educational institution"*(which includes *High school*), *"Research institution but not Educational institution"* (which includes *Academy of sciences*), *"Organization but not Educational institution"*(which includes *Sports organization*) and *"University but not Organization"* will be constructed. After that, SIR will retrieve a certain number of instances of superposition concepts to be annotated from a large-scale text corpus based on the budgets. To equip CE with the ability of extracting universal concepts and SIR with the ability of instance retrieval, we learn CE and SIR on large-scale, easily accessible web resources (Chen et al., 2022), which contains 56M sentences with more than 31K concepts from Wikipedia and Wikidata.

[1]Only minimal additional instances (based on the number of types) will be provided, and therefore, we call our work few-shot NER.

We conduct experiments on 5 few-shot NER benchmarks with different granularity. Experiments show that SuperCD significantly outperforms baselines and other active learning approaches under the same annotation budgets. Furthermore, SuperCD can effectively identify and discriminate superposition concepts. These demonstrate the effectiveness of SuperCD.[2]

Generally speaking, the contributions of this paper are:

- We identify the precise generalization challenge which is a task-intrinsic of few-shot NER and is ignored by previous literature.
- We propose to resolve the precise generalization challenge by discriminating superposition concepts and propose an "A but not B" manner to identify superposition concept set.
- We propose Superposition Concept Discriminator (SuperCD), an active learning framework that injects generalization knowledge to FS-NER models by requiring annotators to provide a minimal number of additional annotated instances of superposition concepts.

2. Related work

Few-shot NER. Previous works of FS-NER focused on better learning strategy and model architecture. Metric-based methods(Snell et al., 2017) are common on many FS-NER benchmarks with different structures like prototype network (Fritzler et al., 2019; Tong et al., 2021; Wang et al., 2022b; Ji et al., 2022; Wang et al., 2022a) and nearest neighbor network (Yang and Katiyar, 2020). Prompt-based methods are promising for FS-NER (Cui et al., 2021; Liu et al., 2022a; Ma et al., 2022b), which fully exploit the knowledge of pre-trained language models. Different learning strategies like contrastive learning(Das et al., 2022), meta learning (Li et al., 2020b,a; de Lichy et al., 2021; Ma et al., 2022c) and self-training (Huang et al., 2021; Wang et al., 2021b) have been used to improve FS-NER. Recently, label semantics like type name and description is proved to be effective for FS-NER (Hou et al., 2020; Wang et al., 2021a; Ma et al., 2022a; Chen et al., 2022; Yang et al., 2022; Lee et al., 2022), but it is difficult to obtain the description or type name accurately. Different from previous works, we focus on the task-intrinsic precise generalization challenge of FS-NER which cannot be addressed by designing better learning strategy or model architecture without introducing more generalization knowledge.

[2]The code this paper: `https://github.com/chen700564/supercd`.

Active learning NER. For NER, most of the previous active learning methods are uncertainty sampling (Shen et al., 2018; Huang et al., 2018; Agrawal et al., 2021; Sapci et al., 2021; Zhou et al., 2021; Liu et al., 2022b), i.e., selecting the instances to be annotated based on the uncertainty scores. In addition, other related work considers data selection by multi-criteria (Shen et al., 2004; Kim, 2020; Nguyen et al., 2022),annotation cost (Wei et al., 2019) etc. Different from previous works, we focus on few-shot scenarios with extremely limited budgets. Furthermore, the proposed SuperCD aims to introduce the generalization knowledge to FS-NER models by annotating minimal instances of superposition concepts for model training, which is different from previous works like uncertainty sampling methods.

3. Superposition Concept Discriminator

As we mentioned above, it is impossible to tackle the precise generalization challenge based only on the knowledge entailing in given few-shot instances. As a result, it is necessary to introduce additional instances to measure superposition concepts. Therefore, the main challenge here is how to identify superposition concepts and produce high-value instances for sufficient supervision with minimum annotation efforts. To this end, we propose Superposition Concept Discriminator (SuperCD), an active learning-based framework for FS-NER. The overall architecture of SuperCD is shown in Figure 2. Specifically, given a few illustrative instances of the novel type, SuperCD first uses a concept extractor to extract the common concepts entailed in the illustrative instances. Then superposition concept sets are constructed based on an "A but not B" manner. After that, a Superposition Instance Retriever is applied to identify instances of each superposition concepts sets from large-scale raw corpus. Finally, annotators are asked to annotate superposition instances according to budget and all annotated and illustrative instances are used to learn few-shot NER models. In the following, we will first describe some critical concepts, and then illustrate each component of SuperCD.

3.1. Definition of Superposition Concept

In this paper, we use the term ***concept*** to refer to a specific generalization of entity mentions, e.g., *city*, *human settlement*, *location*, etc. Note that concepts here are universal, which may be or not be an entity type of a specific FS-NER task. In this paper, We collect 30k concepts from objects of "Instance-of" and "Subclass-of" relations in Wikidata.

We use the term ***superposition concept*** to refer to the concept that is associated with some common concepts entailed in illustrative instances. Without additional information provided, an instance of superposition concept can both be or not be regarded as congener of illustrative instances, and therefore the target entity type can not be clarified unless providing sufficient measurement on these superposition concepts. For the example in Figure 1, *High school* serves as a superposition concept because discriminating them is necessary to determine whether the target type is *University* or *Educational institution*. Unfortunately, due to the large scale of concepts and their compositions, it is very difficult to accurately recognize all superposition concepts for each FS-NER task. But in the following section, we will describe how we obtain the set of superposition concepts using the proposed concept extractor of SuperCD.

Without additional information, it is unclear whether an instance of the superposition concept should be considered as a congener of illustrative instances. Therefore, the target entity type cannot be determined without sufficient measurement on these superposition concepts. In Figure 1, for example, *High school* serves as a superposition concept because distinguishing it is necessary to determine whether the target type is *University* or *Educational institution*. Unfortunately, accurately recognizing all superposition concepts for each FS-NER task can be challenging due to the large scale of concepts and their compositions. However, in the following section we will describe how we obtain the set of superposition concepts in SuperCD.

3.2. Superposition Concept Extraction

As we mentioned above, the main challenge for precise few-shot NER is how to identify the superposition concepts and how to provide additional knowledge about these superposition concepts to FS-NER models.

Unfortunately, it is impractical to directly recognize all superposition concepts due to the large amount of concepts in real world. In this paper, instead of directly identifying each superposition concept, we propose to construct sets of superposition concepts using an elimination-based method. Figure 3 shows an illustration of the entire superposition concept extraction procedure. Specifically, we first introduce a concept extractor, which can generalize illustrative instances into their common concepts. Then the superposition concepts are covered in the different sets of these common concepts.

Formally, given a few illustrative instances $x_1, x_2, ..., x_n$, the concept extractor will first generalize each instance x_i into its corresponding universal concepts using a sequence-to-sequence con-

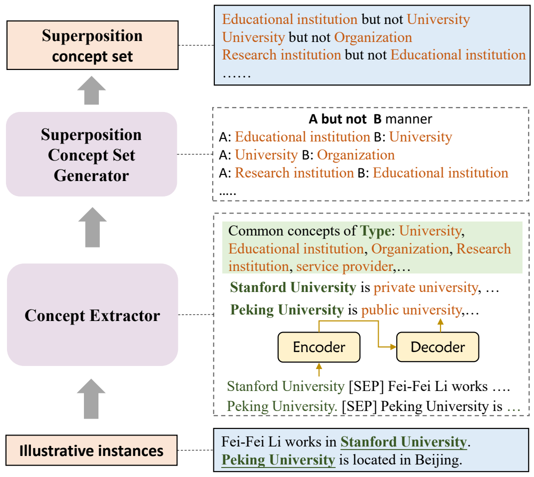


Figure 3: The process of superposition concept extraction. Sets of superposition concepts are constructed through the concepts of few-shot illustrative instances with the "A but not B" manner.

cept extraction model (Chen et al., 2022). The model takes illustrative instances x_i with an annotated entity mention as input, and outputs the corresponding universal concepts:

$$C_i = \mathrm{CE}(x_i) \quad (1)$$

where $C_i = \{c_{i1}, c_{i2}, ..., c_{ik}\}$ is the possible generalized concepts of illustrative instance x_i. Then we collect concepts generated from all illustrative instances, and regards concepts with high appearing frequencies as the common concepts of these few-shot instances. For example, given two illustrative instances in Figure 1, concept extractor will extract *University*, *Educational institution*, *Research institution* as the common concepts. We use $C^* = \{c_1^*, c_2^*, ..., c_m^*\}$ to represents the extracted common concepts.

After that, we obtain sets of superposition concepts using an elimination-based method in an "A but not B" manner. Specifically, sets of superposition concepts are represented as "c_i^* but not c_j^*". For example, given the common concepts {*University*, *Educational institution*, *Research institution*, *Organization*}, superposition concept sets like *"Educational institution but not Organization"*, *"Organization but not University"* and *"Educational institution but not University"* will be constructed, and the superposition concepts *High school*, *Academy of sciences* and *Sports organization* can be covered by the former sets respectively. In this way, we can obtain the superposition concept sets without knowing the exact superposition concepts.

3.3. Superposition Instance Retrieval

Given sets of superposition concepts, another challenge to resolve the precise generalization issue is how to discriminate them and inject the information into FS-NER models. One promising approach is to annotate a minimal number of instances of these superposition concepts, and directly uses them to learn FS-NER models. Therefore, it is necessary to accurately retrieve instances corresponding to superposition concepts from a large-scale corpus.

To this end, we design Superposition Instance Retriever (SIR), a dense retrieval architecture which regards the utterance of sets of superposition concepts as query, and retrieves texts containing such instances from the corpus. Formally, given a text piece x in large-scale corpus and all superposition concept sets of the target entity type, we first construct superposition concept set queries by combining all sets with the same excluded concept. That is, the queries are in the form of $q =$ "$c^*_k | c^*_1, ..., c^*_{k-1}, c^*_{k+1}, ..., c^*_m$", which represents that we want the retrieve instances of concept sets like "c^*_1 but not c^*_k", "c^*_2 but not c^*_k" at the same time. Then the superposition instance retriever will first encode the query and text into dense representations using a deep neural network model respectively:

$$\mathbf{x} = \mathrm{SIR}(x), \mathbf{q} = \mathrm{SIR}(q) \tag{2}$$

After that, the confidence score $s_{x,q}$ indicating the text piece containing a mention of superposition concepts in the set is calculated by:

$$s_{x,q} = \mathbf{x} \circ \mathbf{q} \tag{3}$$

where $\circ$ is inner product. We then iteratively select instances of each query with highest confidence scores as the candidate superposition instances to annotate. If the number of queries larger than the annotation budget, we will order the queries based on the frequencies of the excluded concept.

Finally, annotators are asked to annotate the instances retrieved from SIR. Then additional annotated instances and original illustrative instances are used to learn FS-NER models. In this way, the high-value precise generalization knowledge entailed in the additional annotated instances can be injected to the FS-NER models.

4. Learning CE and SIR

To equip Concept Extractor with ability of extracting universal concepts and Superposition Instance Retriever with ability of instance retrieval, we learn them on large-scale, easily accessible Wikipedia and Wikidata. For CE, we follow SDNet (Chen et al., 2022) to learn the ability to extract concepts from texts. For SIR, we train it through a contrastive learning paradigm.

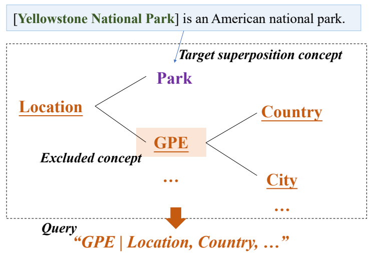


Figure 4: An example of query generation in dataset construction.

4.1. Learning Concept Extractor

Concept extractor is a sequence-to-sequence model which maps the input instances into their corresponding concepts. In this paper, we leverage the dataset constructed by Chen et al. (2022) to train concept extraction. The dataset contains 56M instances in the form of text-to-concepts parallel sequences, which can be directly used to learn the ability of concept extraction.

4.2. Learning Superposition Instance Retriever

To equip SIR with the ability of instance retrieval, we construct large-scale query-text pairs from Wikipedia and Wikidata, and leverage a contrastive learning paradigm to learn the retrieval model. Specifically, we regard all sentences in Wikipedia with anchor words to Wikidata items as training instances. We first randomly sampled a corresponding type from a positive instance in Wikidata as the target superposition concept. Then, the excluded concept in the query is selected from the siblings of target concept, and the remaining concepts in the query are sampled from ancestors and descendants of the excluded concept. For example in Figure 4, given an instance "[Yellowstone National Park] is an American national park." and the target concept *Park* is first sampled, we then random choose a sibling concept *GPE* as the excluded concept and the remaining concepts in the query contain *Location*, *Country*, etc. After that, for each query, we will sample two kinds of negative instances, including instances of the excluded concept and randomly sampled instances does not satisfy the query condition. Finally, the dataset contains 10M query and positive instance pairs, where each pair contains about 200 negative instances.

After obtaining the dataset, we propose to use a contrastive learning paradigm to train SIR. Specifically, given a query q, a positive instance $x^{(+)}$ and several negative instances $X^- = \{x_1^{(-)}, \ldots, x_N^{(-)}\}$, SIR is learned by optimizing the following loss func-

		Vanilla	Random	ALPS	BERT-KM	Vote-k	SuperCD
WNUT17	BERT	27.53	29.06	26.34	28.62	27.92	**34.16**
	StructShot (Yang and Katiyar, 2020)	30.40	33.64	33.57	34.58	33.56	**34.74**
	NSP (Huang et al., 2021)	34.20	43.67	40.15	40.69	39.97	**44.36**
	CONTaiNER (Das et al., 2022)	32.50	35.23	35.23	35.14	34.81	**36.77**
	SDNet (Chen et al., 2022)	44.10	45.07	44.87	44.93	43.03	**45.63**
WNUT16	BERT	26.65	35.04	31.93	30.65	34.34	**37.22**
	StructShot (Yang and Katiyar, 2020)	31.63	32.72	33.15	34.25	33.19	**34.80**
	NSP (Huang et al., 2021)	38.40	42.26	42.06	43.14	42.02	**45.09**
	CONTaiNER (Das et al., 2022)	31.07	32.30	32.53	33.80	32.97	**35.52**
	SDNet (Chen et al., 2022)	47.29	49.89	49.98	50.67	49.68	**50.72**
CoNLL	BERT	67.88	73.86	74.20	73.26	70.54	**74.42**
	StructShot (Yang and Katiyar, 2020)	74.80	76.63	76.74	77.22	**78.75**	77.52
	NSP (Huang et al., 2021)	61.40	76.21	75.25	76.15	74.50	**77.97**
	CONTaiNER (Das et al., 2022)	75.80	78.18	78.76	79.22	80.09	**80.17**
	SDNet (Chen et al., 2022)	71.40	75.88	74.88	74.99	76.15	**76.17**
ACE05	BERT	62.66	65.77	68.19	65.84	65.06	**68.23**
	StructShot (Yang and Katiyar, 2020)	50.63	51.40	51.84	53.37	50.83	**53.73**
	NSP (Huang et al., 2021)	63.73	67.29	66.82	67.10	66.56	**68.23**
	CONTaiNER (Das et al., 2022)	64.12	65.63	66.74	65.04	65.98	**67.68**
	SDNet (Chen et al., 2022)	64.78	69.43	70.99	70.03	69.44	**71.68**
AVE		50.05	53.96	53.71	53.93	53.47	**55.74**

Table 1: Micro-F1 scores of 5+5-shot FS-NER on test set. Vanilla indicates that the model is trained using the initial illustrative data. The annotated budget is 5 sentences for each type. AVE are the average scores of these datasets and few-shot NER models.

tion:

$$L(q, x^+, X^-) = -\log \frac{e^{s_{x^+,q}}}{e^{s_{x^+,q}} + \sum_{i=1}^{N} e^{s_{x_i^-,q}}} \quad (4)$$

5. Experiments

It is important to note that SuperCD does not aim to achieve state-of-the-art results in FS-NER, but rather to solve the intrinsic precise generalization challenge. Previous FS-NER methods can leverage our approach to improve their performance. In addition, SuperCD is a model-agnostic active learning method that can be universally applied to all types of FS-NER models.

5.1. Settings

Active learning setting. Unlike previous active learning settings (Shen et al., 2018; Sapci et al., 2021), this paper employs an extremely limited budget setting that is consistent with the few-shot goal. Specifically, we assume that the budget is related to the number of target type, and at most $M \times N$ sentences can be annotated (N is the number of target entity types). In a K-shot few-shot setting, active FS-NER will be in a K+M-shot configuration. For each FS-NER dataset, we conduct main experiments on 5+5-shot setting. Following Huang et al. (2021), we sample k sentences from the training set to construct initial illustrative instances of each target type in k-shot setting, and the remaining sentences in the training set are used as the unlabeled corpus for active learning. We sample 10 different sets of the illustrative instances. We evaluate the model on the test set with the metric of the average micro-f1 over the 10 runs.

Datasets. We conducted experiment on 4 few-shot NER datasets with different granularity: 1) WNUT17 (Derczynski et al., 2017); 2) ACE2005[3], we use the ACE05-E processed by Wadden et al. (2019) and Lin et al. (2020); 3) CoNLL2003 (Sang and Meulder, 2003); 4) WNUT16 (Strauss et al., 2016). WNUT17, ACE2005, CoNLL focus on coarse type like *location* and *organization*. WNUT16 focus on fine type like *company*.

Baselines. To verify the universality of the proposed method, we conduct experiments on five different FS-NER models in three different types: 1) Linear classifier-based FS-NER models, including BERT (base-uncased) (Devlin et al., 2019) and noising supervised pre-trained RoBERTa (NSP) (Huang et al., 2021); 2) metric-based FS-NER models, including StructShot (Yang and Katiyar, 2020) and CONTaiNER (Das et al., 2022); 3) generative FS-NER model, SDNet (Chen et al., 2022). For StructShot and CONTaiNER, we pretrain BERT using OntoNotes as described in the original papers. Due to the varied architectures, logits-based active learning methods like Entropy (Sapci et al., 2021) cannot be directly applied for all these models. Hence, we primarily evaluate model-agnostic universal active learning methods.

[3] https://www.ldc.upenn.edu/collaborations/past-projects/ace

Additionally, in Section 5.3, we compare SuperCD with logits-based active learning methods using specific FS-NER models.

We compared SuperCD with universal active learning approaches for all FS-NER models: 1) random sampling (**Random**), which selects sentences from unlabeled corpus randomly; 2) Diversity sampling **BERT-KM** (Yuan et al., 2020), which selects diversity sentences based on feature space; 3) **ALPS** (Yuan et al., 2020), which combines the uncertainty and diversity to select sentences. 4) **Vote-k** (SU et al., 2023)[4], which selects diverse, representative instances for annotation.

5.2. Main Results

The experimental results are shown in Table 1. We can see that:

1) **SuperCD can effectively improve the performance of FS-NER by retrieving high-value instances.** Compared with the vanilla models, SuperCD improves the performance by 11.4%, and compared with random sampling baselines, SuperCD improves the performance by 3.3%, which indicates that instances retrieved by SuperCD are high-value and can significantly improve performance on FS-NER.

2) **Discriminating superposition concepts is helpful to resolve precise generalization challenge.** Compared with best active learning baselines, SuperCD improves the performance by 3.4%, which shows that precise generalization challenge is not the problem of identifying boundary or diversity cases which is a concern in traditional active learning methods, and therefore other active learning approaches cannot be used to resolve precise generalization challenge directly.

3) **SuperCD is a universal method for FS-NER.** SuperCD performs well on datasets of varying granularity. Additionally, it is a model-agnostic method that performs well on FS-NER models with different architectures.

	WNUT17		WNUT16		AVE
	BERT	NSP	BERT	NSP	
FT-BERTKM (Yuan et al., 2020)	30.37	39.99	33.26	41.58	36.30
EnTropy (Sapci et al., 2021)	32.04	42.19	**37.34**	45.00	39.14
Badge (Ash et al., 2020)	31.38	44.33	35.14	41.02	37.97
CAL (Margatina et al., 2021)	33.96	39.70	32.23	42.61	37.13
SuperCD	**34.16**	**44.36**	37.22	**45.09**	**40.21**

Table 2: The result of logits-based active learning methods. We conduct 5+5-shot setting.

[4]We conduct the fast vote-k since because it can achieve similar performance to vote-k while being more computationally efficient.

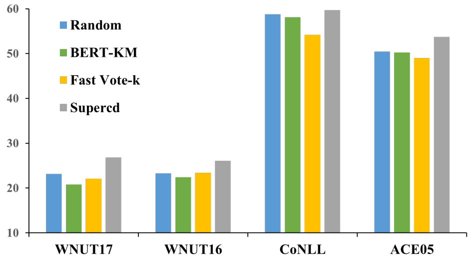


Figure 5: The micro-F1 scores of BERT on entities with unseen concepts in the test set.

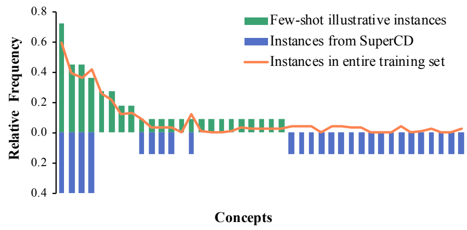
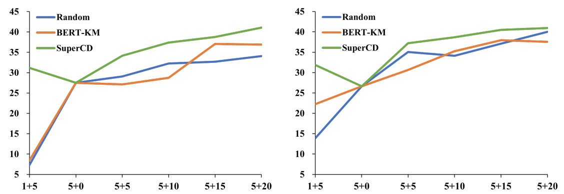


Figure 6: The relative frequency distributions of concepts for few-shot illustrative instances, the annotated instances by SuperCD and entire training set instances in *Location* of WNUT17.



(a) WNUT17. (b) WNUT16.

Figure 7: Performances under different annotated budgets. We use **BERT-KM** to represent the baseline method.

5.3. Comparing SuperCD and logits-based methods.

To further validate the effectiveness of SuperCD, we conduct logits-based active learning methods as additional baselines: including **Entropy** (Sapci et al., 2021), fine-tuned BERT-KM (**FT-BERTKM**) (Yuan et al., 2020), **Badge** (Ash et al., 2020) and **CAL** (Margatina et al., 2021). The result is shown in Table 2. We can see that SuperCD can achieve better or competitive performance across datasets of varying granularity. In addition, we found that for coarse-grained dataset WNUT17, SuperCD provides more significant performance improvements than fine-grained dataset WNUT16. This is because fine-grained datasets suffer from a lower risk of under-generalization and the main problem arises from over-generalization. In contrast, coarse-grained datasets suffer from both over-generalization and under-generalization, and therefore SuperCD achieves a more significant

	WNUT17	
Target types	**Common concepts of illustrative instances**	**Entities in samples from SuperCD**
Location	*City*, *Country*, *Human settlement*, *Educational insititution*, *Location*......	Glenveagh national park → *park* UNF Arena → *arena*
Test text	. . . . places such as <location>**Siachen Glacier**</location> are . . .	
Vanilla	. . . . places such as Siachen Glacier are provided with the	
SuperCD (ours)	. . . . places such as <location>**Siachen Glacier**</location> are . . .	
	WNUT16	
Target types	**Common concepts of illustrative instances**	**Entities in samples vetoed by SuperCD**
company	*Organization*, *Company*, *Website*, *Social networking service*, *Business*,......	NRSC → *government agency* Stanford→ *university*
Test text	FBI hack : ex-special agent says those responsible ...	
Vanilla	<company>**FBI**</company> hack : ex-special agent says those responsible ...	
SuperCD (ours)	FBI hack : ex-special agent says those responsible ...	

Table 3: Cases from WNUT17 and WNUT16. For WNUT17, the vanilla BERT is under-generalized and cannot recognize “Siachen Glacier” as *Location*. For WNUT16, the vanilla BERT is over-generalized and misidentifies “FBI” as *Company*. Note that “Siachen Glacier” and “FBI” are not in the instances retrieved by SuperCD.

improvement.

5.4. Effectiveness of identifying superposition concepts

To further validate the effectiveness of SuperCD, we evaluate the performance of BERT on entities in the test set that contain concepts not seen in the initial illustrative instances. The result is shown in Figure 5. We can see that: 1) Compared to the performance on the full test set, entities with unseen concepts perform significantly worse, which reflects the difficulty of the model to learn accurate generalization of the relevant concept by a few illustrative instances, demonstrating the impact of the precise generalization challenge in FS-NER; 2) SuperCD can alleviate the problem by discriminating the superposition concepts. Compare with baselines, SuperCD can significantly improve the performance on entities with unseen concepts, which verifies that retrieved instances play an important role in the performance improvement of FS-NER.

The above conclusions reveal that SuperCD can effectively identify superposition concepts. We further illustrate this by analyzing the relative frequency distributions of concepts in illustrative instances and annotated instances from SuperCD. We take *location* in WNUT17 as an example and the result is shown in Figure 6. Since the distribution of entire instances is long-tailed, few-shot illustrative instances contain only a few tailed concepts, which makes it difficult in precise generalization. In contrast, instances retrieved by SuperCD contains many tailed concepts which are high-value superposition concepts and cannot be accessed by simply sampling more instances. This demonstrates that SuperCD can effectively identify superposition concepts.

5.5. Effectiveness of SuperCD w.r.t. Annotation Budgets

To further validate the effectiveness of SuperCD, we evaluate the performance of BERT under different budget scenarios, i.e., annotating instances from 5 to 20 for each type and we also conduct 1+5-shot setting. The result is shown in Figure 7, we can see that 1) SuperCD works well under different budget scenarios, which indicates that SuperCD is a promising active learning framework for FS-NER. 2) SuperCD outperforms the baseline approach, which indicates that the instances of superposition concepts help FS-NER models determine what the target entity type is.

6. Case study

We illustrate the effectiveness of SuperCD by some cases shown in Table 3. For coarse-grained dataset WNUT17, when the common concepts of illustrative instances are concepts like *City* and *Country*, vanilla BERT fails to recognize “Siachen Glacier” as *Location*. This indicates that the model is under-generalized and has difficulty in determining whether *Glacier* is part of the target type. In contrast, SuperCD discriminates the superposition concepts by annotating some instances of *Park* and *Arena*. In this way, the FS-NER model can understand the scope of the target type better and avoid under-generalization.

For fine-grained dataset WNUT16, when the target type is *Company*, vanilla BERT misidentifies “FBI” as it. Note that the concepts of “FBI” contain some common concepts of illustrative instances such as *Organization*. This indicates that the model over-generalize *Company* to *Organization*. In contrast, SuperCD retrievals entity mentions of superposition concepts *University* and *Government agency*. These superposition concepts are vetoed as the corresponding entity mentions are not annotated as target types. By learning from additional instances of superposition concepts, the FS-NER model can understand that the target type is not *Organization* and avoid over-generalization.

7. Conclusion

In this paper, we propose Superposition Concept Discriminator for FS-NER which resolves the precise generalization challenge by requiring annotators to annotate minimal additional high-value instances of superposition concepts. We learn the import components of SuperCD – Concept Extractor and Superposition Instance Retriever through large-scale, easily accessible web resources. Experiments on 5 datasets show that SuperCD is effective. For future work, we will extend superposition concepts and SuperCD to other few-shot tasks like object detection and event detection.

8. Limitations

SuperCD is an active learning algorithm, therefore currently human annotators are still needed, although there are very limited instances to annotate. Furthermore, the performance of SuperCD may be influenced by error propagation from the Concept Extractor and the Superposition Instance Retriever. For instance, the concept coverage of Concept Extractor may influence the performance of SuperCD, which can be improved by introducing more knowledge to learn Concept Extractor.

9. Acknowledgments

This work is supported by the Natural Science Foundation of China (No. 62106251, 62122077 and 62306303).

10. Bibliographical References

Ankit Agrawal, Sarsij Tripathi, and Manu Vardhan. 2021. Active learning approach using a modified least confidence sampling strategy for named entity recognition. *Progress in Artificial Intelligence*, 10(2):113–128.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Jiawei Chen, Qing Liu, Hongyu Lin, Xianpei Han, and Le Sun. 2022. 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.

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, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 140–147. 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.

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*, pages 993–1000.

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, ACL 2020, Online, July 5-10, 2020*, pages 1381–1393. Association for Computational Linguistics.

Han Huang, Hongyu Wang, and Dawei Jin. 2018. A low-cost named entity recognition research based on active learning. *Scientific Programming*, 2018.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2021. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, Online and Punta Cana, Dominican Republic. 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.

Yekyung Kim. 2020. Deep active learning for sequence labeling based on diversity and uncertainty in gradient. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 1–8, Suzhou, China. Association for Computational Linguistics.

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.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

Jing Li, Billy Chiu, Shanshan Feng, and Hao Wang. 2020a. Few-shot named entity recognition via meta-learning. *IEEE Transactions on Knowledge and Data Engineering*.

Jing Li, Shuo Shang, and Ling Shao. 2020b. Metaner: Named entity recognition with meta-learning. In *Proceedings of The Web Conference 2020*, pages 429–440.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Andy T Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. 2022a. Qaner: Prompting question answering models for few-shot named entity recognition. *arXiv preprint arXiv:2203.01543*.

Mingyi Liu, Zhiying Tu, Tong Zhang, Tonghua Su, Xiaofei Xu, and Zhongjie Wang. 2022b. Ltp: a new active learning strategy for crf-based named entity recognition. *Neural Processing Letters*, pages 1–22.

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.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minh Van Nguyen, Nghia Ngo, Bonan Min, and Thien Nguyen. 2022. FAMIE: A fast active learning framework for multilingual information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 131–139, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

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, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Ali Osman Berk Sapci, Öznur Tastan, and Reyyan Yeniterzi. 2021. Focusing on possible named entities in active named entity label acquisition. *CoRR*, abs/2111.03837.

Burr Settles. 2009. Active learning literature survey.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 589–596, Barcelona, Spain.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

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*, NIPS'17, page 4080–4090, Red Hook, NY, USA. Curran Associates Inc.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Hongjin SU, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Meihan Tong, Shuai Wang, Bin Xu, Yixin Cao, Minghui Liu, Lei Hou, and Juanzi Li. 2021. Learning from miscellaneous other-class words for few-shot named entity recognition. 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 6236–6247, Online. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. 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 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Jianing Wang, Chengyu Wang, Chuanqi Tan, Minghui Qiu, Songfang Huang, Jun Huang, and Ming Gao. 2022a. SpanProto: A two-stage span-based prototypical network for few-shot named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3476, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peiyi Wang, Runxin Xu, Tianyu Liu, Qingyu Zhou, Yunbo Cao, Baobao Chang, and Zhifang Sui. 2022b. 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.

Yaqing Wang, Haoda Chu, Chao Zhang, and Jing Gao. 2021a. Learning from language description: Low-shot named entity recognition via decomposed framework. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1618–1630, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and

Ahmed Hassan Awadallah. 2021b. Meta self-training for few-shot neural sequence labeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1737–1747.

Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, and Hua Xu. 2019. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322.

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.

Zeng Yang, Linhai Zhang, and Deyu Zhou. 2022. SEE-few: Seed, expand and entail for few-shot named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2540–2550, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Baohang Zhou, Xiangrui Cai, Ying Zhang, Wenya Guo, and Xiaojie Yuan. 2021. Mtaal: multi-task adversarial active learning for medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14586–14593.