Cross-Lingual Contrastive Learning for Fine-Grained Entity Typing for Low-Resource Languages

Xu Han[1], **Yuqi Luo**[1], **Weize Chen**[1], **Zhiyuan Liu**[1,2,3,4*]
Maosong Sun[1,2,3,4*], **Botong Zhou**[5], **Fei Hao**[5], **Suncong Zheng**[5]
[1] Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University
Beijing National Research Center for Information Science and Technology
[2] Institute Guo Qiang, Tsinghua University
[3] International Innovation Center of Tsinghua University
[4] Beijing Academy of Artificial Intelligence, BAAI
[5] Tencent AI Platform Department, Tencent Inc
{hanxu17,yq-luo19,chenwz21}@mails.tsinghua.edu.cn
{liuzy,sms}@tsinghua.edu.cn

Abstract

Fine-grained entity typing (FGET) aims to classify named entity mentions into fine-grained entity types, which is meaningful for entity-related NLP tasks. For FGET, a key challenge is the low-resource problem — the complex entity type hierarchy makes it difficult to manually label data. Especially for those languages other than English, human-labeled data is extremely scarce. In this paper, we propose a cross-lingual contrastive learning framework to learn FGET models for low-resource languages. Specifically, we use multi-lingual pre-trained language models (PLMs) as the backbone to transfer the typing knowledge from high-resource languages (such as English) to low-resource languages (such as Chinese). Furthermore, we introduce entity-pair-oriented heuristic rules as well as machine translation to obtain cross-lingual distantly-supervised data, and apply cross-lingual contrastive learning on the distantly-supervised data to enhance the backbone PLMs. Experimental results show that by applying our framework, we can easily learn effective FGET models for low-resource languages, even without any language-specific human-labeled data. Our code is also available at https://github.com/thunlp/CrossET.

1 Introduction

Recently, various efforts have been devoted to exploring fine-grained entity typing (FGET) (Ling and Weld, 2012; Li et al., 2020), aiming to identify concrete fine-grained entity types for named entity mentions in sentences (Figure 1). Since the type information of named entity mentions is useful for understanding textual semantics, FGET is widely applied to enhance entity-related tasks, such as coreference resolution (Khosla and Rose, 2020), entity linking (Onoe and Durrett, 2020; Chen et al., 2020a), relation extraction (Ren et al., 2017; Zhou and Chen, 2021) and event extraction (Nguyen et al., 2016; Yang and Mitchell, 2016).

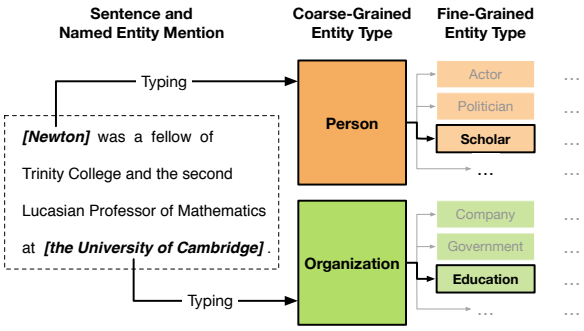

Figure 1: The illustration of classifying named entity mentions in sentences into fine-grained entity types.

Despite the success of FGET, the low-resource problem is always a challenge of FGET, since the complex type hierarchy makes it difficult to manually label data. To alleviate the low-resource problem, besides utilizing auto-labeled data (Ling and Weld, 2012; Gillick et al., 2014; Xin et al., 2018; Dai et al., 2021), manually building FGET datasets is the most effective approach (Sang and De Meulder, 2003; Hovy et al., 2006; Ling and Weld, 2012; Choi et al., 2018; Ding et al., 2021). However, existing FGET datasets are mainly in English. For datasets in specific languages other than English, such as Chinese (Lee et al., 2020), Japanese (Suzuki et al., 2016), Dutch and Spanish (van Erp and Vossen, 2017), their scale and quality are not comparable to those English datasets. In this paper, we introduce a cross-lingual framework to learn FGET models for low-resource languages, via utilizing the data in high-resource languages

*Corresponding authors.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2241 - 2250

(e.g. utilizing English datasets).

Transferring the typing knowledge from high-resource languages to low-resource languages is not easy. As different languages have quite different patterns, it is challenging to understand the semantics of both high-resource and low-resource languages at the same time. With only a few examples of low-resource languages and no parallel data, it is also hard to bridge different languages for knowledge transfer. To handle these issues: (1) we use multi-lingual pre-trained language models (PLMs) as backbone. Multi-lingual PLMs such as M-BERT (Devlin et al., 2019) are pre-trained on large-scale multi-lingual corpora, taking it as the backbone can well encode data in different languages into the same semantic space (Han et al., 2021). (2) we apply heuristic rules and cross-lingual contrastive learning to bridge multiple languages. We design several entity-pair-oriented heuristic rules to obtain distant supervision, which can automatically annotate entity types by utilizing latent relations between entity pairs. Machine translation is used on the auto-labeled data to establish a connection between high-resource and low-resource languages. Finally, we apply contrastive learning to learn similarities between cross-lingual auto-labeled types, instead of using pseudo-labels to learn a classifier, which can enhance the type recognition ability and reduce the side effect of auto-labeled data.

For convenience, we name our cross-lingual contrastive learning framework "CROSS-C" in the following sections. We conduct experiments on two popular FGET datasets: Open-Entity (Choi et al., 2018) and Few-NERD (Ding et al., 2021), and translate their test sets into non-English versions to evaluate the effectiveness of CROSS-C for low-resource languages. Quantitative experimental results show that applying CROSS-C can easily train effective FGET models for low-resource languages, even without any language-specific human-labeled data. Besides quantitative experiments, we also provide some visualization of feature spaces and conduct case studies for qualitative analysis to show how CROSS-C works.

2 Method

In this section, we will introduce our cross-lingual framework to learn FGET models for low-resource languages. We will first give some essential notations and definitions, and then elaborate on the details of our framework.

2.1 Notations and Definitions

As shown in Figure 1, given a sentence x and one named entity mention m in the sentence, our goal is to determine types from a fine-grained type set $\mathcal{T}$ according to the sentence context for the mention m. Note that FGET is a multi-label classification problem, since multiple types can be assigned to a single named entity mention.

For a high-resource language h, sufficient human-labeled data $\{\mathcal{X}_h, \mathcal{Y}_h\}$ exists, where $\mathcal{X}_h = \{x_{h,1}, x_{h,2}, \ldots\}$ is the sentence set and $\mathcal{Y}_h = \{y_{h,1}, y_{h,2}, \ldots\}$ is the label set. Each sentence $x_{h,i} \in \mathcal{X}_h$ contains a named entity mention $m_{h,i}$, and $y_{h,i} \subseteq \mathcal{T}$ is the fine-grained type set of the named entity mention $m_{h,i}$.

Similarly, we define the dataset $\{\mathcal{X}_l, \mathcal{Y}_l\}$ for a low-resource language l, where $|\mathcal{X}_l| \ll |\mathcal{X}_h|$ [1]. In this paper, we use $\{\mathcal{X}_h, \mathcal{Y}_h\}$, $\{\mathcal{X}_l, \mathcal{Y}_l\}$ and large-scale unlabeled multi-lingual data to train a FGET model for the low-resource language l.

2.2 Multi-Lingual Pre-Trained Encoder

We use multi-lingual BERT (M-BERT) (Devlin et al., 2019) as the framework backbone to encode the input. M-BERT has the same architecture as BERT, but is pre-trained on the multi-lingual corpora in 104 languages. Therefore, M-BERT has a good ability to transfer knowledge across languages (Pires et al., 2019; Selvaraj et al., 2021), making it suits our setting well. Note that, our framework does not depend on a specific PLM, any other multi-lingual PLMs can also be used as the backbone to encode the input.

Given a sentence $x = [w_1, \ldots, m, \ldots, w_n]$, where m is the named entity mention, we additionally insert an entity marker `[ENT]` on each side of the mention m. By feeding the sentence with entity markers into M-BERT, we can get representations $[\mathbf{h}_{w_1}, \ldots, \mathbf{h}_{\texttt{[ENT]}}, \mathbf{h}_m, \mathbf{h}_{\texttt{[ENT]}}, \ldots, \mathbf{h}_{w_n}]$ for all input tokens. The left entity marker representation $\mathbf{h}_{\texttt{[ENT]}}$ is used to represent the named entity mention. For simplicity, we denote this process as $\mathbf{m} = \texttt{M-PLM}(x)$, where $\mathbf{m}$ is the entity mention representation and x is the input sentence. Given each entity type $t \in \mathcal{T}$, the probability that the mention m in the sentence x can be classified as

[1] In our experiments, we focus on handling a difficult and extreme case $|\mathcal{X}_l| = 0$, i.e, there is no any human-labeled data for the low-resource language l.

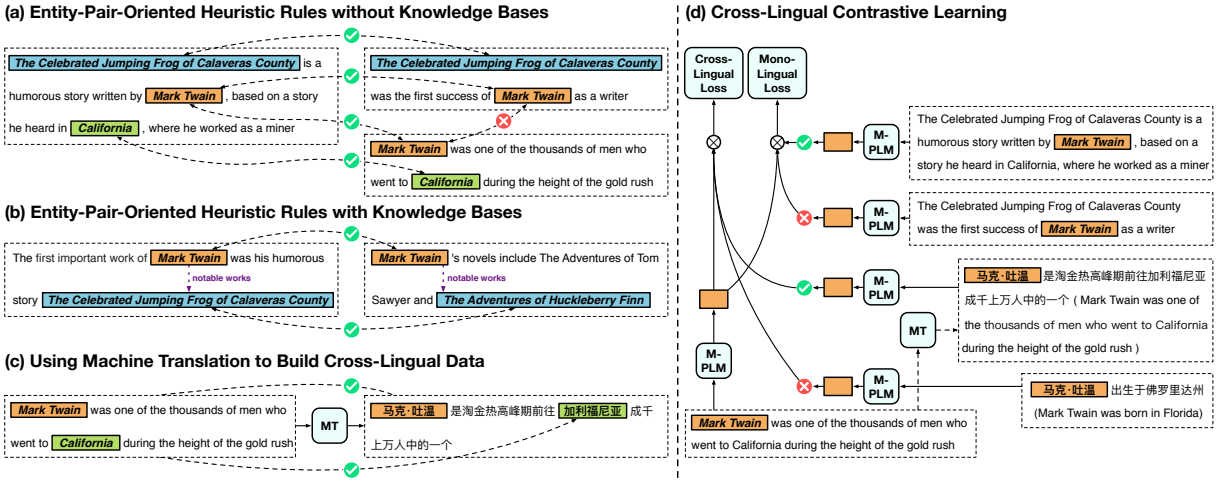

Figure 2: The illustration of entity-pair-oriented heuristic rules and cross-lingual contrastive learning. In this example, we want to transfer the typing knowledge from English to Chinese.

the type t is given as

$$P_\theta(t|x) = \sigma\big(\mathbf{t}^{\intercal}\texttt{M-PLM}(x)\big), \tag{1}$$

where σ is the sigmoid function, $\mathbf{t}$ is the representation of the entity type t, and θ indicates all learnable model parameters.

With the data $\{\mathcal{X}_h, \mathcal{Y}_h\}$ in the high-resource language h and the data $\{\mathcal{X}_l, \mathcal{Y}_l\}$ in the low-resource language l, the overall optimization objective is as

$$\arg\max_\theta \Big[\mathcal{L}_{\text{high}}(\theta) + \mathcal{L}_{\text{low}}(\theta)\Big], \tag{2}$$

where $\mathcal{L}_{\text{high}}(\theta)$ and $\mathcal{L}_{\text{low}}(\theta)$ respectively indicate the loss functions for the high-resource language h and the low-resource language l. These loss functions are defined as

$$\begin{aligned}\mathcal{L}_{\text{high}}(\theta) = \frac{1}{|\mathcal{X}_h|}\sum_{i=1}^{|\mathcal{X}_h|}\sum_{t\in\mathcal{T}}\big[&\delta_{t\in y_{h,i}}\log P_\theta(t|x_{h,i}) \\ &+ (1-\delta_{t\in y_{h,i}})\log(1-P_\theta(t|x_{h,i}))\big], \\ \mathcal{L}_{\text{low}}(\theta) = \frac{1}{|\mathcal{X}_l|}\sum_{i=1}^{|\mathcal{X}_l|}\sum_{t\in\mathcal{T}}\big[&\delta_{t\in y_{l,i}}\log P_\theta(t|x_{l,i}) \\ &+ (1-\delta_{t\in y_{l,i}})\log(1-P_\theta(t|x_{l,i}))\big].\end{aligned} \tag{3}$$

For the function δ_c, if the condition c is satisfied, then $\delta_c = 1$, otherwise $\delta_c = 0$.

2.3 Heuristic Rules for Data Augmentation

As we mentioned before, there are only a few human-labeled examples in low-resource languages. Although multi-lingual PLMs can provide an effective backbone to understand multilingual semantics, more examples are still required to bridge different languages.

The existing distantly-supervised methods annotate the mentions of the same entity in multiple sentences with the same pseudo label (Ling and Weld, 2012; Gillick et al., 2014; Xin et al., 2018). However, in Figure 2, the mention "Mark Twain" requires to be annotated with "writer" or "miner" according to specific semantics. Hence, these single-entity-oriented heuristic rules inevitably bring much noise.

To this end, we introduce heuristic rules orienting entity pairs to automatically annotate data with less noise. Instead of annotating specific entity types, we annotate whether two named entity mentions are of similar types. On the one hand, this strategy can consider the correlation and similarity between different types. On the other hand, this strategy is suitable for contrastive learning, which can reduce the side effect of data noise. In fact, in relation extraction, recent works have adopted similar strategies (Soares et al., 2019; Peng et al., 2020) and achieved promising results. More specifically, as shown in Figure 2, we take three rules to obtain distantly-supervised data:

(1) **Rules without knowledge bases**. As shown in Figure 2(a), without using knowledge bases, if one entity pair is mentioned by two sentences, the mentions of the same entity in these two sentences are considered to have similar types.

(2) **Rules with knowledge bases**. As shown in Figure 2(b), by using knowledge bases, if entity pairs in two sentences have same relations in knowledge bases, and these pairs have shared entities, the mentions of corresponding entities are considered to have similar types.

(3) **Building cross-lingual data with machine translation**. As shown in Figure 2(c), we use machine translation to translate the data from the high-resource language to the low-resource language. Owing to the translation, the above-mentioned auto-labeled examples and their translated versions constitute a cross-lingual distantly-supervised dataset.

By making full advantage of distant supervision and machine translation, we can greatly expand our dataset to bridge high-resource and low-resource languages, and further transfer the typing knowledge between these languages. To make FGET models pay more attention to textual contexts rather than merely focusing on entity names, we use the `[MASK]` token to mask named entity mentions with a probability of 0.5.

2.4 Cross-Lingual Contrastive Learning

With all above-mentioned heuristic rules in Section 2.3, we can get the distantly-supervised data $\tilde{\mathcal{X}}_h = \{\tilde{x}_{h,1}, \tilde{x}_{h,2}, \ldots\}$ in the high-resource language h, the distantly-supervised data $\tilde{\mathcal{X}}_l = \{\tilde{x}_{l,1}, \tilde{x}_{l,2}, \ldots\}$ in the low-resource language l, and the translated data $\tilde{\mathcal{X}}_t = \{\tilde{x}_{t,1}, \tilde{x}_{t,2}, \ldots\}$. Given any two sentences x_1, x_2 in these distantly-supervised datasets, we use the function $s(x_1, x_2)$ to measure the similarity between the entity mentions of the two sentences. In practice, we take the cosine similarity with temperature τ as the function $s(x_1, x_2)$:

$$s(x_1, x_2) = \frac{\texttt{M-PLM}(x_1)^\intercal \texttt{M-PLM}(x_2)}{\|\texttt{M-PLM}(x_1)\| \cdot \|\texttt{M-PLM}(x_2)\| \cdot \tau}, \quad (4)$$

where $\texttt{M-PLM}(\cdot)$ is the entity mention representation computed by multi-lingual PLMs.

The cross-lingual contrastive learning consists of two important objectives. One is the mono-lingual objective for each language, and the other is the cross-lingual objective. For both the high-resource language h and the low-resource language l, their mono-lingual objectives are defined as follows,

$$\begin{aligned}\mathcal{L}_{\text{mono-h}}(\theta) = \frac{1}{|\tilde{\mathcal{X}}_h|}\sum_{i=1}^{|\tilde{\mathcal{X}}_h|}\big[\log \sum_{\tilde{p}\in\mathcal{P}(\tilde{x}_{h,i})} e^{s(\tilde{x}_{h,i},\tilde{p})} \\ - \log(\sum_{\tilde{p}\in\mathcal{P}(\tilde{x}_{h,i})} e^{s(\tilde{x}_{h,i},\tilde{p})} + \sum_{\tilde{n}\in\mathcal{N}(\tilde{x}_{h,i})} e^{s(\tilde{x}_{h,i},\tilde{n})})\big], \\ \mathcal{L}_{\text{mono-l}}(\theta) = \frac{1}{|\tilde{\mathcal{X}}_l|}\sum_{i=1}^{|\tilde{\mathcal{X}}_l|}\big[\log \sum_{\tilde{p}\in\mathcal{P}(\tilde{x}_{l,i})} e^{s(\tilde{x}_{l,i},\tilde{p})} \\ - \log(\sum_{\tilde{p}\in\mathcal{P}(\tilde{x}_{l,i})} e^{s(\tilde{x}_{l,i},\tilde{p})} + \sum_{\tilde{n}\in\mathcal{N}(\tilde{x}_{l,i})} e^{s(\tilde{x}_{l,i},\tilde{n})})\big],\end{aligned} \quad (5)$$

where $\mathcal{P}(\tilde{x}_{h,i}) \subseteq \tilde{\mathcal{X}}_h$ and $\mathcal{N}(\tilde{x}_{h,i}) \subseteq \tilde{\mathcal{X}}_h$ are respectively the positive set and the negative set of the example $\tilde{x}_{h,i}$. $\mathcal{P}(\tilde{x}_{l,i})$ and $\mathcal{N}(\tilde{x}_{l,i})$ are defined in a similar way for the example $\tilde{x}_{l,i}$.

To ensure that the model does not push the representations of different languages far away, so that the low-resource language l can benefit from the high-resource language h, we further use $\tilde{\mathcal{X}}_h$ and its translated set $\tilde{\mathcal{X}}_t$ to define the cross-lingual objective as follows,

$$\begin{aligned}\mathcal{L}_{\text{cross}}(\theta) = \frac{1}{|\tilde{\mathcal{X}}_t|}\sum_{i=1}^{|\tilde{\mathcal{X}}_t|}\big[\log \sum_{\tilde{p}\in\mathcal{P}(\tilde{x}_{t,i})} e^{s(\tilde{x}_{t,i},\tilde{p})} \\ - \log(\sum_{j=1}^{|\tilde{\mathcal{X}}_t|}\delta_{i\neq j} e^{s(\tilde{x}_{t,i},\tilde{x}_{t,j})} + \sum_{j=1}^{|\tilde{\mathcal{X}}_h|} e^{s(\tilde{x}_{t,i},\tilde{x}_{h,j})})\big],\end{aligned} \quad (6)$$

where $\mathcal{P}(\tilde{x}_{t,i}) \subseteq \tilde{\mathcal{X}}_h \cup \tilde{\mathcal{X}}_t$ is the positive set of the example $\tilde{x}_{t,i}$. The final objective of the cross-lingual contrastive learning is to optimize

$$\arg\max_{\theta} \left[\mathcal{L}_{\text{cross}}(\theta) + \mathcal{L}_{\text{mono-r}}(\theta) + \mathcal{L}_{\text{mono-l}}(\theta)\right]. \quad (7)$$

2.5 Pre-Training and Fine-Tuning

We divide the whole learning process into two stages: pre-training and fine-tuning. The pre-training stage is to use Eq. (7) to optimize parameters on the distantly-supervised data. Considering computational efficiency, every time we sample a batch of examples for contrastive learning, and then sample multiple positive examples for each example in the batch. After the pre-training stage, we use Eq. (2) to fine-tune parameters on human-labeled data to learn classifiers for FGET.

3 Experiment

In this section, we evaluate the effectiveness of our framework CROSS-C on two typical entity-related datasets: Open-Entity and Few-NERD. For each dataset, we conduct experiments in both low-resource (few-shot or zero-shot) and full-set settings. In addition to quantitative experiments, to further show how our method works, we also provide some visualization of feature spaces for qualitative analysis.

3.1 Dataset Settings

Open-Entity (Choi et al., 2018) and **Few-NERD** (Ding et al., 2021) are both popular FGET datasets. Open-Entity includes 9 general types and 121 fine-grained types. Each example in Open-Entity may correspond to multiple entity types.

Few-NERD includes 8 general types and 66 fine-grained types. Both of these two datasets have a clear type hierarchy, which is suitable for evaluating the model performance on the entity typing task. In our experiments, we require models to predict both general types and fine-grained types for each entity mention in sentences.

3.2 Experimental Settings

In this paper, we select English as a high-resource language and Chinese as a low-resource language. We attempt to use human-labeled English data and large-scale unlabeled multi-lingual data for learning, to obtain an effective Chinese FGET model. This is very difficult, since no any Chinese human-labeled data is used in this process.

To obtain distantly-supervised data, we apply our heuristic rules to automatically annotate the English and Chinese Wikipedia pages [2]. We then use machine translation (Klein et al., 2017; Tan et al., 2020) to translate the English distant-supervised examples into corresponding Chinese versions for cross-lingual contrastive learning.

All test sets of Open-Entity and Few-NERD are translated into Chinese for evaluation. Although the test set built by machine translation may exist some errors, the overall semantics of the translated examples can still support determining the types of entity mentions. Taking human-labeled examples for evaluation is better, yet large-scale human-annotated entity typing datasets are still lacking. The experiments are performed under three settings:

Few-shot setting. This setting requires models to infer entity types with a few supervised examples. We randomly sample 2, 4, 8, 16 examples for each entity type for training.

Zero-shot setting. This setting requires models to infer entity types without any supervised training, i.e., no human-labeled example is used for training.

Full-set setting. In this setting, all supervised examples in datasets are used for training.

We follow the widely-used setting of Ling and Weld (2012), use the loose micro F_1 scores to evaluate the performance of models.

3.3 Baseline Settings

We use M-BERT (Devlin et al., 2019) as the backbone [3] to implement all baseline models and our model **"CROSS-C"**. We use **"F-T"** to denote directly using English human-labeled data to fine-tune M-BERT, which is demonstrated the effectiveness in Selvaraj et al. (2021). We use **"MONO-C"** to denote only using mono-lingual contrastive learning objectives for pre-training, and then use English human-labeled data to fine-tune pre-trained parameters. All above-mentioned models are optimized by AdamW with the learning rate {5e-6,1e-5,3e-5,5e-5}. The batch size used for pre-training and fine-tuning is from {8,16,32,64,128,256}. For cross-lingual contrastive learning, we only traverse large-scale distantly-supervised data once. For fine-tuning models on human-labeled data, the epochs are from {1,3,5,7,10}. The temperature τ used for the cosine similarity is 0.5.

3.4 The Overall Performance in Low-Resource Settings

The results of few-shot entity typing for Chinese are reported in Table 1. The table shows that:

(1) Using a multi-lingual PLM as the backbone can lead to an effective FGET model for those low-resource languages. All methods, including both the baseline models and our CROSS-C, can achieve non-trivial entity typing results on the Chinese test sets, without using any Chinese human-labeled examples for training models.

(2) Using distantly-supervised data for contrastive learning can significantly improve the typing capabilities of the backbone PLMs. Compared with directly fine-tuning a multi-lingual PLM with human-labeled data in high-resource languages, conducting contrastive learning on multi-lingual distantly-supervised data can better bridge high-resource languages and low-resource languages, which is beneficial to obtain effective models in low-resource languages.

(3) Compared with mono-lingual contrastive learning, our cross-lingual contrastive learning can better improve the transfer of typing knowledge from high-resource languages to low-resource languages. Our CROSS-C achieves the best results in all shot settings. And the improvements of CROSS-C will gradually increase as the number of shots decreases. These results show that our method can effectively improve model performance for low-resource languages even without any high-quality supervised language-specific data.

We also report the entity typing performance on the original English test sets in Table 2. From the

[2] https://dumps.wikimedia.org/
[3] https://github.com/google-research/bert

Dataset	Model	2-Shot			4-Shot			8-Shot			16-Shot		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Open-Entity	F-T	71.4	20.8	32.2	69.4	26.4	38.3	71.3	30.3	42.5	71.3	45.8	55.8
	MONO-C	48.9	38.6	43.1$^{\uparrow 10.9}$	51.2	45.7	48.3$^{\uparrow 10.0}$	56.7	49.8	53.1$^{\uparrow 10.6}$	63.5	58.6	60.9$^{\uparrow 5.1}$
	CROSS-C	56.4	42.0	**48.1**$^{\uparrow 15.9}$	58.3	43.8	**50.1**$^{\uparrow 11.8}$	60.7	51.1	**55.5**$^{\uparrow 13.0}$	70.2	59.9	**64.6**$^{\uparrow 8.8}$
Few-NERD	F-T	72.7	25.1	37.3	73.2	35.7	48.0	71.8	44.1	54.7	69.2	51.7	59.2
	MONO-C	54.2	41.7	47.2$^{\uparrow 9.9}$	64.3	51.2	57.0$^{\uparrow 9.0}$	65.9	56.4	60.8$^{\uparrow 6.1}$	67.8	60.4	63.9$^{\uparrow 4.7}$
	CROSS-C	56.4	45.7	**50.5**$^{\uparrow 13.2}$	66.3	56.3	**60.9**$^{\uparrow 12.9}$	70.3	62.4	**66.1**$^{\uparrow 11.4}$	69.9	66.0	**67.9**$^{\uparrow 8.7}$

Table 1: The model performance (%) on the Chinese test sets. All these models are learned in the few-shot learning setting. K-Shot means that each entity type has only K examples for training. $\uparrow$ represents the improvements over F-T.

Dataset	Model	2-Shot			4-Shot			8-Shot			16-Shot		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Open-Entity	F-T	69.2	35.7	47.1	67.5	43.1	52.6	68.7	49.6	57.6	65.9	55.5	60.3
	MONO-C	59.1	44.3	50.6$^{\uparrow 3.5}$	57.8	50.2	53.7$^{\uparrow 1.1}$	59.6	56.3	57.9$^{\uparrow 0.3}$	61.1	60.9	**61.0**$^{\uparrow 0.7}$
	CROSS-C	56.8	45.9	**50.8**$^{\uparrow 3.7}$	61.0	51.7	**55.9**$^{\uparrow 3.3}$	61.6	58.2	**59.8**$^{\uparrow 2.2}$	59.5	62.1	60.8$^{\uparrow 0.5}$
Few-NERD	F-T	78.4	38.3	51.4	79.2	49.6	61.0	80.0	61.4	69.5	78.4	67.9	72.8
	MONO-C	56.3	48.3	52.0$^{\uparrow 0.6}$	63.0	61.4	62.2$^{\uparrow 1.2}$	76.9	70.8	**73.7**$^{\uparrow 4.2}$	78.7	70.2	74.2$^{\uparrow 1.4}$
	CROSS-C	66.6	52.0	**58.4**$^{\uparrow 7.0}$	73.8	63.5	**68.3**$^{\uparrow 7.3}$	75.9	69.3	72.4$^{\uparrow 2.9}$	78.8	73.1	**75.9**$^{\uparrow 3.1}$

Table 2: The model performance (%) on the English test sets. All these models are learned in the few-shot learning setting. K-Shot means that each entity type has only K examples for training. $\uparrow$ represents the improvements over F-T.

Model	Open-Entity (Chinese)		
	P	R	F_1
M-BERT	8.8	4.0	5.5
CROSS-C	**24.3**$^{\uparrow 15.5}$	**11.4**$^{\uparrow 15.3}$	**15.5**$^{\uparrow 10.0}$
Model	Open-Entity (Chinese)		
	P	R	F_1
M-BERT	6.2	3.5	4.5
CROSS-C	**25.5**$^{\uparrow 19.3}$	**13.5**$^{\uparrow 10.0}$	**17.7**$^{\uparrow 13.2}$

Table 3: The zero-shot performance (%) on the Chinese test sets. All these models do not use any supervised examples to tune models for entity typing. $\uparrow$ represents the improvements over M-BERT.

table we can see:

(1) In our low-resource settings, although there are no human-labeled Chinese data at all, there are still some high-quality English examples for each entity type. Therefore, the improvements of contrastive learning on the English test sets are not as obvious as on the Chinese test sets. However, compared with directly fine-tuning PLMs, contrastive learning methods still bring significant improvements, demonstrating the power of using distant supervision for data augmentation.

(2) Owing to multi-lingual data, which makes models in multiple languages learn from each other, our cross-lingual contrastive learning further brings additional improvements over the mono-lingual contrastive learning. This proves the effectiveness of our cross-lingual contrastive framework.

Table 3 shows the results of zero-shot entity typing on the Chinese test sets. In this table, we can see that: without a trained type classifier, our cross-lingual contrastive learning still brings the backbone PLM a strong type recognition ability in the pre-training stage.

3.5 The Overall Performance in Full-Set Settings

We show the model performance curve as the number of supervised examples increases in Figure 3. Note that only the supervised examples of the high-resource language English are used for training models. There is still no human-labeled data for the low-resource language Chinese. The results in the figure show that:

(1) For high-resource languages, by using more supervised examples, the improvements brought by contrastive learning are gradually decreasing, which is in line with our intuition. But we should also notice that even in the full-set setting, contrastive learning methods achieve comparable or even slightly better results than fine-tuning PLMs. This means that taking contrastive learning can well reduce the impact of data noise while enhanc-

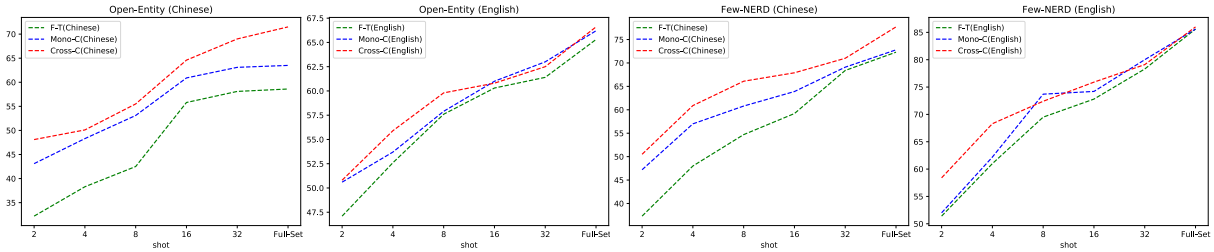

Figure 3: The model performance (%) curve as the number of supervised examples increases. We report the F_1 scores (%) on both the Chinese and English test sets.

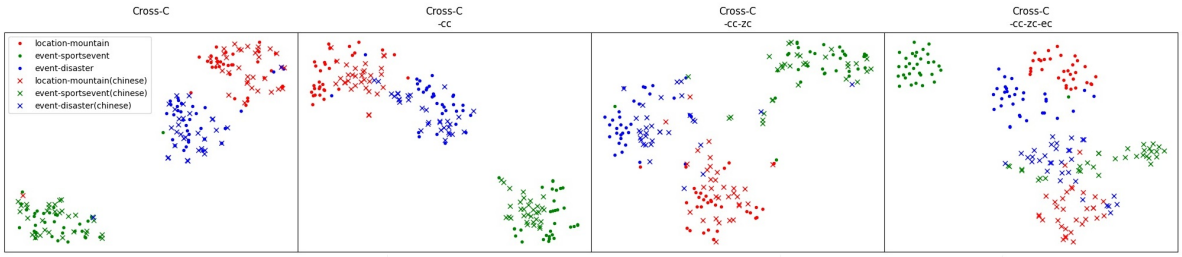

Figure 4: The model visualization during the ablation experiments for CROSS-C. We select some typical entity types in the dataset Few-NERD and their corresponding examples for visualization.

ing performance by making full use of distantly-supervised data.

(2) In both low-resource and full-set settings, the results of our contrastive learning on the Chinese test sets are always significantly higher than other baseline models. This shows that our framework can utilize the supervised data of high-resource languages and large-scale unlabeled multi-lingual data to handle FGET for low-resource languages.

3.6 Ablation Experiments and Model Visualization

In order to show how our CROSS-C works more intuitively, we conduct comprehensive ablation experiments. The results of the ablation experiments are shown in Table , where "−cc" means that we drop the cross-lingual contrastive objective for pre-training the backbone PLM, "−zc" means that we drop the mono-lingual contrastive objective on the Chinese distantly-supervised data, and "−ec" means that we drop the mono-lingual contrastive objective on the English distantly-supervised data. From Table , we can find that: both the mono-lingual contrastive objectives and the cross-lingual objective play an important role in enhancing the backbone PLM, and the combination of them can lead to greater improvements. This is also the reason that our cross-lingual contrastive learning includes both mono-lingual and cross-lingual contrastive objectives for pre-training the backbone.

We also give the visualization of the model during the ablation experiments of CROSS-C in Figure 4. From the visualization results, we can find that it is difficult to bridge high-resource languages and low-resource languages without using any contrastive learning. As we gradually increase the number of contrastive learning objectives, the distinction between entity types becomes more obvious, and the fusion of multi-lingual semantics also becomes better.

4 Related Work

As one of the most important tasks in the field of information extraction, FGET has been studied for a long time. Ling and Weld (2012); Yosef et al. (2012) first propose to classify named entity mentions into various fine-grained entity types, instead of just a few coarse-grained types (Sang and De Meulder, 2003; Hovy et al., 2006). Since fine-grained types bring informative semantics for language understanding, these types are widely used to enhance entity-related NLP tasks, such as coreference resolution (Khosla and Rose, 2020), entity linking (Onoe and Durrett, 2020; Chen et al., 2020a), relation extraction (Ren et al., 2017; Zhou and Chen, 2021) and event extraction (Nguyen et al., 2016; Yang and Mitchell, 2016). Some recent efforts further incorporate entity types to learn

Model	Open-Entity (Chinese)			
	2-Shot	4-Shot	8-Shot	16-Shot
CROSS-C	**48.1**	**50.1**	**55.5**	**64.6**
−cc	$45.2^{\downarrow 2.9}$	$46.9^{\downarrow 3.2}$	$48.1^{\downarrow 7.4}$	$55.7^{\downarrow 8.9}$
−cc−zc	$43.1^{\downarrow 5.0}$	$48.3^{\downarrow 1.8}$	$53.1^{\downarrow 2.4}$	$60.9^{\downarrow 3.7}$
−cc−zc−ec	$32.2^{\downarrow 15.9}$	$38.3^{\downarrow 11.8}$	$42.5^{\downarrow 13.0}$	$55.8^{\downarrow 8.8}$
Model	Few-NERD (Chinese)			
	2-Shot	4-Shot	8-Shot	16-Shot
CROSS-C	**50.5**	**60.9**	**66.1**	**67.9**
−cc	$48.1^{\downarrow 2.4}$	$59.2^{\downarrow 1.7}$	$63.2^{\downarrow 2.9}$	$64.9^{\downarrow 3.0}$
−cc−zc	$47.2^{\downarrow 3.3}$	$57.0^{\downarrow 3.0}$	$60.8^{\downarrow 5.3}$	$63.9^{\downarrow 4.0}$
−cc−zc−ec	$37.3^{\downarrow 13.2}$	$48.0^{\downarrow 12.9}$	$54.7^{\downarrow 11.4}$	$59.2^{\downarrow 8.7}$

Table 4: The ablation experiments for CROSS-C. We directly report the F_1 scores (%) in the few-shot settings. ↓ represents the amount of the decrease in model performance compared to CROSS-C after giving up some contrastive learning objectives.

entity-enhanced PLMs (Zhang et al., 2019; Sun et al., 2019; Liu et al., 2020).

Distantly-supervised FGET methods. Since entity types have complex hierarchies, manually annotating FGET data is not easy, and thus the low-resource problem is one of the key challenges of FGET. To alleviate this issue, distantly-supervised methods have been widely explored for FGET. One typical distantly-supervised approach is using knowledge bases to automatically annotate entities mentioned in the text. Ling and Weld (2012); Gillick et al. (2014) collect anchors in the Wikipedia pages that correspond to entities in knowledge bases, and then label these anchors with entity types in knowledge bases. This approach is then followed by a series of works (Ren et al., 2017; Xin et al., 2018; Choi et al., 2018; Dai et al., 2019; Jin et al., 2019; Lee et al., 2020) to obtain pseudo labels. Other approaches use various noun phrases in sentences as type pseudo labels (Dai et al., 2020, 2021), which can make full use of the recently proposed PLMs for data augmentation.

Human-labeled FGET datasets. In addition to the distantly-supervised methods, the construction of FGET datasets is also advancing. CoNLL (Sang and De Meulder, 2003) and Ontonotes (Hovy et al., 2006) are the earliest datasets, although they just cover several coarse-grained types. Then, Ling and Weld (2012); Gillick et al. (2014); Ding et al. (2021) introduce about a hundred fine-grained types and annotate a large number of examples for each type. Choi et al. (2018) further extend FGET by introducing an ultra-fine set containing thousands of types. Since annotating FGET examples is time-consuming and labor-intensive, many of the ultra-fine types proposed by Choi et al. (2018) only have distantly-supervised examples. However, all these efforts only focus on English. There are also some efforts to build datasets in other languages, such as Chinese (Lee et al., 2020), Japanese (Suzuki et al., 2016), Dutch and Spanish (van Erp and Vossen, 2017), but the scale and quality of these non-English datasets are still not comparable with English datasets, i.e., non-English human-labeled data are still scarce.

Cross-lingual and contrastive learning for FGET. Although cross-lingual learning has been widely explored in entity linking (Sil et al., 2018; Upadhyay et al., 2018; Rijhwani et al., 2019) and named entity recognition (Pan et al., 2017; Xie et al., 2018; Rahimi et al., 2019; Zhou et al., 2019), cross-lingual entity typing has not yet been explored much (Selvaraj et al., 2021). For contrastive learning (Chen et al., 2020b; Oord et al., 2018), some preliminary works have explored it for extracting relations between entities (Soares et al., 2019) and achieved promising results. Peng et al. (2020) further use contrastive learning to analyze the impact of entity information on relation extraction. Similar to cross-lingual learning, the exploration of contrastive learning for FGET is still in the preliminary stage.

5 Conclusion and Future Work

In this paper, to learn effective FGET models for those low-resource languages, we propose an effective cross-lingual contrastive learning framework CROSS-C to transfer the typing knowledge from high-resource languages to low-resource languages. Specifically, the framework CROSS-C uses a multi-lingual PLM — M-BERT as the framework backbone, which can simultaneously capture multi-lingual semantics in a unified feature space. Furthermore, to bridge the gap between high-resource languages and low-resource languages, we introduce entity-pair-oriented heuristic rules as well as machine translation to automatically obtain high-quality cross-lingual data, and then apply cross-lingual contrastive learning on this distantly-supervised data to enhance the backbone PLM. The experimental results show that by applying CROSS-C, the typing knowledge can be transferred from high-resource languages to low-resource languages, and we can learn effective FGET models without any language-specific

human-labeled data for those low-resource languages. In the future:

(1) We will explore how to better utilize unsupervised data to deal with the low-resource problem of FGET, such as using better PLMs and more effective tuning methods.

(2) We will also promote the construction of cross-lingual FGET datasets, which will advance the development of FGET in specific languages, especially for those low-resource languages other than English.

Acknowledgements

This work is supported by the National Key R&D Program of China (No. 2020AAA0106502), Institute Guo Qiang of Tsinghua University, Beijing Academy of Artificial Intelligence (BAAI), and International Innovation Center of Tsinghua University. This work is also supported by Tencent AI Platform Department.

References

Shuang Chen, Jinpeng Wang, Feng Jiang, and Chin-Yew Lin. 2020a. Improving entity linking by modeling latent entity type information. In *Proceedings of AAAI*, pages 7529–7537.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pages 1597–1607.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of ACL*, pages 87–96.

Hongliang Dai, Donghong Du, Xin Li, and Yangqiu Song. 2019. Improving fine-grained entity typing with entity linking. In *Proceedings of EMNLP-IJCNLP*, pages 6210–6215.

Hongliang Dai, Yangqiu Song, and Xin Li. 2020. Exploiting semantic relations for fine-grained entity typing. In *Proceedings of AKBC*.

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-fine entity typing with weak supervision from a masked language model. *arXiv preprint arXiv:2106.04098*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Hai-Tao Zheng, and Zhiyuan Liu. 2021. Few-nerd: A few-shot named entity recognition dataset. In *Proceedings of ACL*, pages 3198–3213.

Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh. 2014. Context-dependent fine-grained entity type tagging. *arXiv preprint arXiv:1412.1820*.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of NAACL-HLT*, pages 57–60.

Hailong Jin, Lei Hou, Juanzi Li, and Tiansi Dong. 2019. Fine-grained entity typing via hierarchical multi graph convolutional networks. In *Proceedings of EMNLP-IJCNLP*, pages 4969–4978.

Sopan Khosla and Carolyn Rose. 2020. Using type information to improve entity coreference resolution. In *Proceedings of the Workshop of ACL*, pages 20–31.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL, System Demonstrations*, pages 67–72.

Chin Lee, Hongliang Dai, Yangqiu Song, and Xin Li. 2020. A chinese corpus for fine-grained entity typing. In *Proceedings of LREC*, pages 4451–4457.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *TKDE*.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Proceedings of AAAI*, pages 94–100.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of AAAI*, pages 2901–2908.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL-HLT*, pages 300–309.

Yasumasa Onoe and Greg Durrett. 2020. Fine-grained entity typing for domain independent entity linking. In *Proceedings of AAAI*, pages 8576–8583.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of ACL*, pages 1946–1958.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of EMNLP*, pages 3661–3672.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of ACL*, pages 4996–5001.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. In *Proceedings of ACL*, pages 151–164.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of WWW*, pages 1015–1024.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of AAAI*, pages 6924–6931.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of NAACL-HLT*, pages 142–147.

Nila Selvaraj, Yasumasa Onoe, and Greg Durrett. 2021. Cross-lingual fine-grained entity typing. *arXiv preprint arXiv:2110.07837*.

Avirup Sil, Gourab Kundu, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual entity linking. In *Proceedings of AAAI*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of ACL*, pages 2895–2905.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. 2016. Fine-grained named entity classification with wikipedia article vectors. In *Proceedings of WI*, pages 483–486.

Zhixing Tan, Jiacheng Zhang, Xuancheng Huang, Gang Chen, Shuo Wang, Maosong Sun, Huanbo Luan, and Yang Liu. 2020. Thumt: an open-source toolkit for neural machine translation. In *Proceedings of AMTA*, pages 116–122.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of EMNLP*, pages 2486–2495.

Marieke van Erp and Piek Vossen. 2017. Multilingual fine-grained entity typing. In *Proceedings of LDK*, pages 262–275.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime G Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of EMNLP*, pages 369–379.

Ji Xin, Hao Zhu, Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Put it back: Entity typing with language model enhancement. In *Proceedings of EMNLP*, pages 993–998.

Bishan Yang and Tom Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of HAACL-HLT*, pages 289–299.

Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol, and Gerhard Weikum. 2012. Hyena: Hierarchical type classification for entity names. In *Proceedings of COLING*, pages 1361–1370.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of ACL*, pages 1441–1451.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of ACL*, pages 3461–3471.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *arXiv preprint arXiv:2102.01373*.