Cross-Lingual Name Entity Recognition from Clinical Text using Mixed Language Query

Kunli Shi[1], Gongchi Chen[1], Jinghang Gu[2], Longhua Qian[1*], and Guodong Zhou[1]

[1] School of Computer Science and Technology, Soochow University, Suzhou, China
[2] Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hong Kong, China
1098592026@163.com
20224227020@stu.suda.edu.cn
gujinghangnlp@gmail.com
qianlonghua@suda.edu.cn
gdzhou@suda.edu.cn

Abstract. Cross-lingual Named Entity Recognition (Cross-Lingual NER) addresses the challenge of NER with limited annotated data in low-resource languages by transferring knowledge from high-resource languages. Particularly, in the clinical domain, the lack of annotated corpora for Cross-Lingual NER hinders the development of cross-lingual clinical text named entity recognition. By leveraging the English clinical text corpus I2B2 2010 and the Chinese clinical text corpus CCKS2019, we construct a cross-lingual clinical text named entity recognition corpus (CLC-NER) via label alignment. Further, we propose a machine reading comprehension framework for Cross-Lingual NER using mixed language queries to enhance model transfer capabilities. We conduct comprehensive experiments on the CLC-NER corpus, and the results demonstrate the superiority of our approach over other systems.

Keywords: Cross-Lingual NER, Clinical Text, Mixed Language Query, Machine Reading Comprehension.

1 Introduction

Named Entity Recognition (NER) is a task aimed at accurately locating entities within a given text and categorizing them into predefined entity types. It plays a crucial role in many downstream applications such as relation extraction and question answering. The development of deep learning technology has led to significant breakthroughs in this task. However, supervised learning methods often require a large amount of manually annotated training data, which can be costly and time-consuming, especially for low-resource languages. Therefore, many researchers have focused on zero-shot cross-lingual NER scenarios, which involve using annotated data from a resource-rich source language to perform NER in a target language without labeled data.

Zero-shot cross-lingual NER methods can typically be categorized into two types: annotation projection and direct model transfer. Annotation projection utilize annotated data from a source language to generate pseudo-labeled data in the target language[1–3]. Subsequently, they train NER models on the target language, enabling NER in the target language. One drawback of these methods is that the automatic translation used

to generate target language data may introduce translation errors and label alignment errors. On the other hand, direct model transfer learn language-agnostic features through feature space alignment, thereby transferring models trained on the source language to the target language[3–5]. The limitation of these methods is that they require a certain degree of similarity between the source and target languages, making them less applicable to languages with significant differences, such as Chinese and English.

Currently, most research efforts are concentrated on zero-shot cross-lingual NER tasks in general domains, with relatively little exploration in domain-specific cross-lingual NER. For instance, in the field of clinical medicine, the lack of relevant cross-lingual annotated data has hindered the development of cross-lingual clinical name entity recognition tasks. Furthermore, cross-lingual clinical name entity recognition poses more challenges due to variations in data volume, quality, structure, format, as well as differences in the naming conventions, abbreviations, and terminology usage of biomedical entities among different languages.

To facilitate the development of cross-lingual clinical entity recognition, we constructed a corpus for cross-lingual clinical name entity recognition(CLC-NER) using existing monolingual annotated corpora through label alignment. We employed cross-lingual pre-trained models(XLM-R) for knowledge transfer. Additionally, we introduced a machine reading comprehension framework and, based on this, proposed a cross-lingual named entity method using mixed language queries. By integrating prior knowledge from labels in different languages and exploring potential relationships between different annotated corpora in cross-lingual scenarios, we aimed to enhance task transfer performance.

2 Related Work

2.1 Cross-lingual NER corpus

Currently, cross-lingual named entity recognition tasks primarily rely on the CoNLL2002/2003 shared task data[6, 7] and the WikiAnn dataset[8]. The CoNLL2002/2003 dataset includes four closely related languages: English, German, Spanish, and Dutch, and focuses on four types of named entities in the news domain: person (PER), location (LOC), organization (ORG), and miscellaneou (MISC). WikiAnn, on the other hand, is a dataset encompassing 282 languages and includes various entity types such as person(PER), location(LOC), and organization (ORG). Previous research mainly employed CoNLL2002/2003 for studying transfer tasks in languages with similar linguistic systems, while WikiAnn was utilized to evaluate NER transfer performance when dealing with languages with more significant linguistic differences.

2.2 Cross-lingual NER

Based on the shared content between the source language and the target language, cross-lingual named entity recognition methods are typically categorized into two approaches: annotation projection and direct model transfer.

Annotation projection method involves projecting annotated data from the source language to generate pseudo-labeled data in the target language. Previous methods often relied on parallel corpora[9]. Mayhew et al.[1] used a dictionary-based greedy decoding algorithm to establish word-to-word mappings between the source and target languages, reducing the dependency of annotation projection methods on parallel texts. However, word-to-word projection methods cannot consider contextual meaning, which can affect the quality of entity label projection. Jain et al.[10] employed machine translation to translate sentences and entities separately. They used dictionaries to generate candidate matches for translated entities and employed features such as orthography and phonetic recognition to match the translated entities, resulting in high-quality entity annotation projection.

Direct model transfer methods leverage shared representations between two languages, applying a model trained on the source language to the target language. Tsai et al.[4] generated Wikipedia features for cross-lingual transfer by linking the target language to Wikipedia entries. Ni et al.[9] built mapping functions between word vectors in different languages using dictionaries, enabling the mapping of target language vectors into source language vectors. However, direct model transfer cannot utilize lexicalized features when applied to the target language. Therefore, Xie et al.[2] improved upon methods like Ni et al. by incorporating a nearest-neighbor word vector translation approach, effectively leveraging lexicalized features and enhancing model transfer performance.

With the advancement of pre-trained models, models like BERT[11] have made significant progress in natural language understanding tasks by leveraging large-scale unlabeled text corpora for self-supervised learning to acquire latent knowledge in natural language texts. Multilingual models such as mBERT and XLM[9] further propelled the latest developments in cross-lingual understanding tasks. These cross-lingual models are trained on extensive multilingual unlabeled data, obtaining multilingual word embeddings and shared model parameters, thus enabling effective cross-lingual transfer on multilingual corpora. Keung et al.[5] built upon mBERT by using adversarial learning to align word vectors across different languages to enhance task performance. Wu et al.[12] proposed the Teacher-Student Learning (TSL) model for NER task transfer, which involves training a teacher model using source language annotated data and distilling knowledge from the teacher model to a student model using unannotated data in the target language, improving both single-source and multi-source transfer capabilities. Wu et al.[13] introduced the UniTrans framework, employing ensemble learning to fully utilize pseudo-labeled and unlabeled data for knowledge transfer, enhancing data reliability in transfer learning. Li et al. (2022)[14] extended the teacher-student model by proposing a multi-teacher multi-task framework (MTMT). By introducing a similarity task, they trained two teacher models to obtain pseudo-labeled data in the target language, and conducted multi-task learning on the student model, ultimately achieving strong performance on datasets like CoNLL2002/2003.

2.3 Machine reading comprehension

Machine Reading Comprehension (MRC) is originally a natural language understanding task used to test a machine's ability to answer questions given context. Levy et al.[15] were among the first to simplify relation extraction as a reading comprehension problem and effectively extended it to Zero-Shot scenarios. With the rise of deep learning and large-scale datasets, especially after the emergence of pre-trained models like BERT, many MRC systems based on pre-trained models have performed well on question-answering datasets such as SQuAD[16] and MS MARCO[17]. Some researchers began to recognize the versatility of the machine reading comprehension framework. Li et al.[18] proposed applying the MRC framework to named entity recognition, designing specific question templates for different entity categories, and providing a unified paradigm for nested and non-nested entities. To enhance information interaction between entity heads and tails, Cao et al.[19] introduced double affine transformations into MRC, achieving an F1 score of 92.8 on the CCKS2017 dataset. Zheng et al.[20] integrated the CRF-MT-Adapt model and MRC model using a voting strategy, achieving superior performance on the CCKS2020 dataset.

3 Dataset Construction

Due to the lack of existing cross-lingual clinical text Named Entity Recognition (NER) task datasets, we developed a dataset for investigating cross-lingual clinical text NER, referred to as CLC-NER, by aligning the labels of the CCKS 2019 dataset, which is designed for Chinese electronic medical records NER, and the 2010 I2B2/VA dataset, intended for English concept extraction. This alignment process enabled us to unify the labels of the two datasets, forming the basis for our research in cross-lingual clinical text NER.

3.1 CCKS 2019

CCKS 2019(referred to as CCKS)[21] is part of a series of evaluations conducted by CCKS in the context of semantic understanding of Chinese electronic medical records. Building upon the medical named entity recognition evaluation tasks of CCKS2017 and 2018, CCKS2019 extends and expands the scope. It consists of two sub-tasks: medical named entity recognition and medical entity attribute extraction. Our work focuses on the first sub-task, which involves extracting relevant entities from medical clinical texts and identifying them into six predefined categories. These categories include diseases and diagnosis(疾病和诊断), imaging examination(影像检查), laboratory test(实验室检验), surgery(手术), medication(药物), and anatomical site(解剖部位).

3.2 I2B2 2010

The I2B2 2010 dataset[22](referred to as I2B2)was jointly provided by I2B2 and the VA. This evaluation task consists of three sub-tasks: concept extraction, assertion classification, and relation classification. All three sub-tasks share the same dataset, comprising 349 training documents, 477 test documents, and 877 unlabeled documents. However, only a portion of the data has been publicly released after the evaluation. The publicly available I2B2 dataset includes 170 training documents and 256 test documents. Our focus is on the concept extraction task, which defines three concept entity types: medical problem, medical treatment, and medical test.

3.3 Correlation

The above subsection provide descriptions of the concepts or entity types in the two datasets. We can observe that while their annotation schemes differ somewhat, there are certain corresponding relationships between some types. One notable difference is that CCKS includes the "anatomical site" class of entities, used to specify the anatomical site in the human body where diseases or symptoms occur, whereas I2B2 does not annotate such entities.

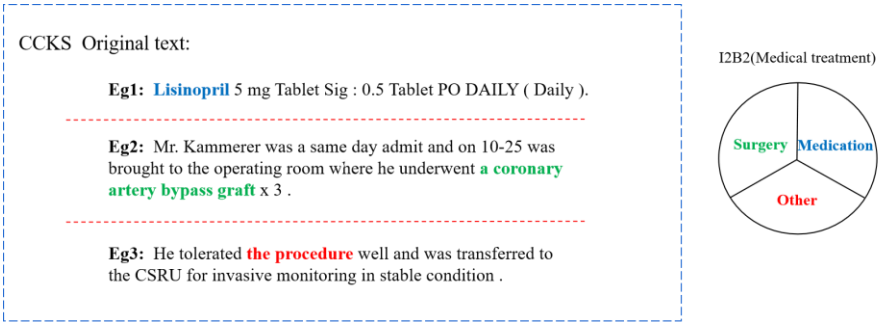

Fig. 1. Differences in entity annotation scope.

On the other hand, the concepts annotated in the I2B2 dataset are broader in scope than the entities in the CCKS dataset. As shown in Figure 1, the "Medical Treatment" type in I2B2 encompasses not only explicit treatment methods, such as "Surgery" and "Medication", as seen in the first two examples, but also includes some general treatment concepts, as in the third example where "the procedure" refers to a certain treatment process. As illustrated in Figure 2, although both "Medical problem" in I2B2 and "Disease and Diagnosis" in CCKS annotate disease names, their scope and granularity differ. "Medical Problem" covers a wider range, including some clinical symptoms, such as infection, redness, and drainage. In contrast, "Disease and Diagnosis" entities strictly adhere to the medical definition of diseases and include fine-grained annotations such as "Hepatoblastoma, hypodermic type (fetal and embryonic) " within the broader category.

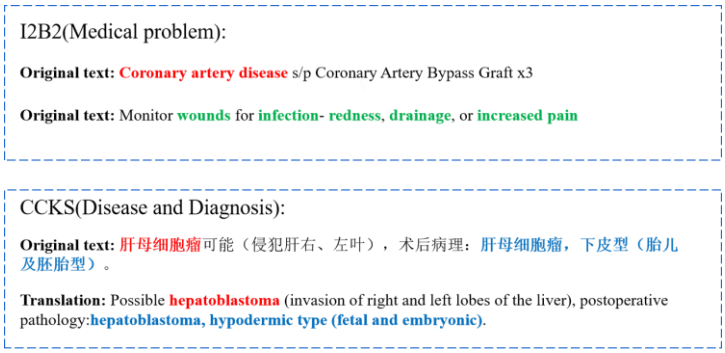

Fig. 2. Differences in annotation scope between "Disease and Diagnosis" and "Medical Problem".

3.4 Label alignment

Based on the similarities and differences between the two corpora, we used a label alignment approach to unify similar concept entity types and discarded entity types that couldn't be aligned. Specifically, we mapped the six entity types in the CCKS dataset to three entity types, aligning them with the annotation scheme of the I2B2dataset. This alignment is shown in Table 1:

Table 1. Label alignment rules between CCKS and I2B2.

CCKS	I2B2	CLC-NER
疾病和诊断(diseases and diagnosis)	Medical problem	Medical problem
影像检查(imaging examination)	Medical test	Medical test
实验室检验(laboratory test)		
手术(surgery)	Medical treatment	Medical treatment
药物(medication)		
~~解剖部位(anatomical site)~~	-	-

From the table, it can be seen that the "Imaging Examination" and "Laboratory Test" entity types in CCKS are similar in meaning to the "Medical Test" concept type in the 2010 I2B2 corpus. Therefore, we grouped "Imaging Examination" and "Laboratory Test" into one category. Similarly, we mapped the "Surgery" and "Medication" entity types in CCKS to the "Medical Treatment" concept type in I2B2. Since there is no corresponding concept type for "Anatomical Site" in the 2010 I2B2 corpus, we removed it.

4 Framework

4.1 Machine reading comprehension

Figure 3 depicts a cross-lingual named entity recognition framework based on the MRC architecture, consisting of three main components: the input layer, encoding layer, and classification layer. Due to the pointer-labeling scheme used for output, multiple questions are posed to the context to extract entities of different types. First, we convert the token sequence generated by concatenating the query and context into vectors through embedding. Next, they are encoded into hidden representations using the XLM-R model. Finally, a classifier determines whether each token marks the beginning or end of entity.

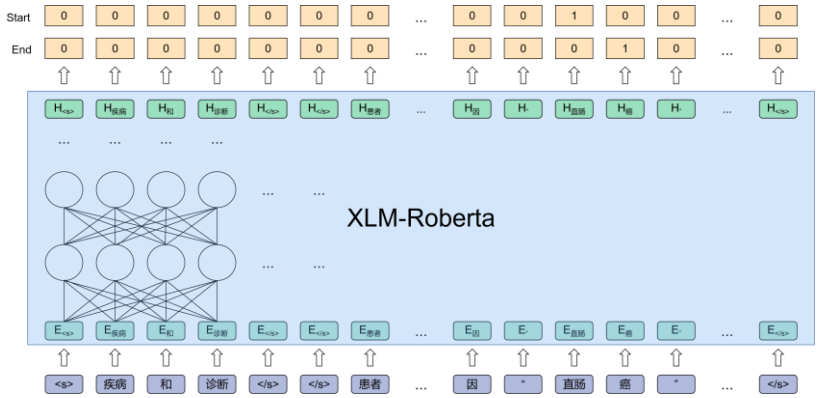

Fig. 3. NER framework based on machine reading comprehension.

Input Layer：
Its role is to segment the text composed of queries and context into token sequences and then transform them into vector sequences through token embedding. Specifically, given an input sequence $X = \{x_i\}_{i=1}^{N}$ with N tokens, it produces a sequence of vectors $V=\{v_i\}_{i=1}^{N}$. $v_i$is the vector corresponding to the i-th token.

Encoding Layer：
The encoder maps the sequence of lexical element vectors from the input layer to a sequence of hidden vectors $H=\{h_i\}_{i=1}^{N}$：

$$H=\text{Encoder}(V) \tag{1}$$

The Encoder model can be any encoder model that uses cross-lingual, in this paper we have chosen the XLM-roberta_base model. h_i is the hidden vector corresponding to the i-th token.

Classification Layer：
After obtaining the hidden vectors for each token, they are fed into the two linear classification layers and the probability distributions for each token as the start and end of the entity are computed using the softmax function, respectively:

$$\mathrm{p}^{\mathrm{s/e}}(x_i)=\mathrm{softmax}(\mathrm{W}^{\mathrm{s/e}}h_i+\mathrm{b}^{\mathrm{s/e}}) \tag{2}$$

$$\hat{\mathrm{y}}^{\mathrm{s/e}}=\mathrm{argmax}(\mathrm{p}^{\mathrm{s/e}}(x_i)) \tag{3}$$

Here $\mathrm{p}^{\mathrm{s}}(x_i)$ and$\mathrm{p}^{\mathrm{e}}(x_i)$ denote the probability that the ith token starts and ends as an entity, respectively, and $\hat{\mathrm{y}}_i^{\mathrm{s/e}}$ denotes the final classification result that the i-th token starts and ends as an entity.

Loss Function：
We use the cross-entropy loss function to compute the loss for the training task, which consists of two components：

$$\mathrm{L}=\mathrm{L}_{\mathrm{START}}+\mathrm{L}_{\mathrm{END}} \tag{4}$$

$$\mathrm{L}_{\mathrm{START}}=\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}-[\mathrm{y}_i^{\mathrm{s}}\log\mathrm{p}_i^{\mathrm{s}}+(1-\mathrm{y}_i^{\mathrm{s}})\log(1-\mathrm{p}_i^{\mathrm{s}})] \tag{5}$$

$$\mathrm{L}_{\mathrm{END}}=\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}-[\mathrm{y}_i^{\mathrm{e}}\log\mathrm{p}_i^{\mathrm{e}}+(1-\mathrm{y}_i^{\mathrm{e}})\log(1-\mathrm{p}_i^{\mathrm{e}})] \tag{6}$$

where $\mathrm{L}_{\mathrm{START}}$ and $\mathrm{L}_{\mathrm{END}}$ are computed as follows, and $\mathrm{y}_i^{\mathrm{s/e}}$ denotes the i-th token's as the real label of the start and end of the entity.

Finally, we use a proximity matching strategy on the final classification result to determine the boundary of an entity.

4.2 Construction of mixed language query

In the context of named entity recognition (NER) based on the machine reading comprehension (MRC) framework, the choice of queries has a notable impact on recognition performance. Similarly, constructing rational and effective queries is highly significant for knowledge transfer in cross-lingual NER.

In monolingual NER, incorporating prior knowledge containing entity type information can induce the model to enhance task performance. However, in the context of cross-lingual NER, which involves multiple languages, using a single query clearly cannot effectively guide the model to learn the prior knowledge across different languages, thus limiting the performance of model transfer. Therefore, this paper proposes a mixed language query construction method, wherein by integrating prior knowledge from multiple languages into the queries, the model can learn the corresponding relationships between different languages, thereby improving the transfer performance of cross-lingual tasks.

Specifically, given a two-language query set $\mathrm{Q}=\{\mathrm{Q}_{\mathrm{zh}},\mathrm{Q}_{\mathrm{en}}\}$, where each language query set contains a priori knowledge of m entity types, i.e:

$$Q_{zh} = \{E_{zh}^{1}, E_{zh}^{2} \ldots E_{zh}^{m}\} \tag{7}$$

$$Q_{en} = \{E_{en}^{1}, E_{en}^{2} \ldots E_{en}^{m}\} \tag{8}$$

where E^{i} denotes the label of the i-th entity type, andE_{zh}^{i} and E_{en}^{i} are translations of each other.

We use the separator "/" to splice the type information of Chinese and English, so as to merge the a priori knowledge of the two languages. As an example, we show the concatenation method with English followed by Chinese, i.e.,:

$$Q_{mix} = \{E_{en}^{1}/E_{zh}^{1}, E_{en}^{2}/E_{zh}^{2} \ldots E_{en}^{m}/E_{zh}^{m}\} \tag{9}$$

4.3 Query template set

In order to investigate the impact of different query templates on model transfer performance, we defined various query templates by combining language and task aspects. This task comprises two language types, Zh (Chinese) and En (English), and two tasks, Src (source task) and Tgt (target task). Taking CCKS as the source task and I2B2 as the target task, we provide an example of the templates conbined from the "Medical treatment" entity type in the CLC-NER corpus, as shown in Table 2.

Table 2. Combination of query templates.

Query Type	Query Templates
Src_Zh	<s>药物、手术</s>
Src_En	<s>medication、surgery</s>
Src_ZhEn	<s>药物/medication、手术/surgery</s>
Src_EnZh	<s>medication/药物、surgery/手术</s>
Tgt_Zh	<s>医疗治疗</s>
Tgt_En	<s>Medical treatment</s>
Tgt_ZhEn	<s>药物/medication、手术/surgery</s>
Tgt_EnZh	<s>medication/药物、surgery/手术</s>

For example, Src_Zh denotes the use of Chinese labels (i.e., "药物" and "手术") from the source language task to generate query templates as prior knowledge. Src_ZhEn represents the generation of mixed language query templates using labels from the source language task, with Chinese first and English second. Similarly, Tgt_ZhEn uses labels from the target language task to create mixed language query templates.

5 Experiments

5.1 Experiment settings

Datasets

The experiments use the CLC-NER introduced in Section 3, and the dataset sizes are shown in Table 3. Both datasets are divided into two subsets for training and testing. "Abstract/Note" and "Entity" denote the number of abstracts and entities in the subset, respectively. It should be noted that the number of entities in the CCKS dataset is the number after excluding the "anatomical site" entities. From the table, we can see that the entity size of I2B2 is larger than that of CCKS, and the entity size of its test set is larger than training set.

Table 3. CLC-NER dataset statistics.

Dataset	Subset	Abstract/Note	Entity
I2B2 2010(En)	Train	170	16,525
	Test	256	31,161
CCKS 2019(Zh)	Train	1,001	9,257
	Test	379	2,908

The number of CLC-NER entities is shown in Table 4, from which it can be seen that the number of entities for "medical problem" is the highest in both corpora. In the I2B2 dataset, there is not much difference between the number of "Medical treatment" entities and the number of "Medical test" entities. In the CCKS dataset, the training set exhibits the lowest count of "Medical treatment" entities, while the test set displays the lowest count of "Medical test" entities.

Table 4. Statistics on the number of entities in the CLC-NER dataset.

Entity Type	I2B2 2010(En)				CCKS 2019(Zh)			
	Training		Test		Training		Test	
	Entity	%	Entity	%	Entity	%	Entity	%
Medical problem	7,073	43	12,592	40	4,242	46	1,323	45
Medical treatment	4,844	29	9,344	30	2,164	23	938	32
Medical test	4,608	28	9,225	30	2,851	31	647	23
Sum	16,525	100	31,161	100	9,257	100	2,908	100

Implementation details

The XLM-R-base trained by Conneau et al.[23] is used as Encoding model. The hyperparameters used for training are listed in Table 5. Throughout this study, all experiments are conducted on a 2080Ti. The standard P/R/F1 metrics are adopted to evaluate the performance.

Table 5. Hyper-Parameter Settings

Hyper Parameter	Value
Batch size	64
Maximum sequence length	128
Learning rate	2e-5
Epoch	10
Dropout	0.1
Optimizer	AdamW

5.2 Experimental results

The impact of different query templates on cross-language transfer performance. Tables 6 and 7 compare the effects of different query templates on the transfer performance in two transfer directions, where the transfer direction in Table 6 is from CCKS source task to I2B2 target task and vice versa in Table 7. Tables (a) and (b) indicate the performance of using the source and target task labels as the query templates. For example, the cell value in the "Src_Zh" row and "Src_En" column indicate the performance when predicting with the Chinese label of the source task on the training set and the English label of the source task on the test set of the target task. Since preliminary experiments show poor performance when different task labels are used for training and testing, this paper only considers the transfer performance between source task labels (four templates) and target task labels (four templates). The experiments take the average of five runs as the final performance value, and the values in the right bracket are the standard variance of the five runs. The same query template was used for training, and the highest performance values for the test templates are shown in bold.

Table 6. The impact of different query templates on cross-language transfer performance(CCKS to I2B2).

(a) the source task (CCKS) labels as query templates

Train\Test	Src_Zh	Src_En	Src_ZhEn	Src_EnZh
Src_Zh	36.5(±3.1)	26.7(±4.2)	35.6(±2.2)	**37.0**(±2.6)
Src_En	22.9(±4.7)	**38.1**(±1.5)	34.7(±2.2)	37.7(±1.6)
Src_ZhEn	30.0(±1.7)	31.9(±2.3)	**38.7**(±1.0)	37.2(±1.2)
Src_EnZh	31.1(±3.0)	33.8(±3.0)	38.1(±0.9)	**38.7**(±0.6)

(b) the target task (I2B2) labels as query templates

Train\Test	Tgt_Zh	Tgt_En	Tgt_ZhEn	Tgt_EnZh
Tgt_Zh	38.7(±1.6)	35.1(±5.2)	**38.8**(±1.4)	38.7(±1.5)
Tgt_En	30.6(±5.2)	**39.7**(±1.7)	37.9(±2.6)	39.7(±1.9)
Tgt_ZhEn	35.1(±2.3)	35.3(±2.9)	**40.8**(±1.6)	39.9(±1.6)
Tgt_EnZh	37.0(±1.8)	38.9(±1.1)	39.1(±1.8)	**39.7**(±1.3)

Table 7. The impact of different query templates on cross-language transfer performance(I2B2 to CCKS).

(a) the source task (I2B2) labels as query templates

Train\Test	Src_Zh	Src_En	Src_ZhEn	Src_EnZh
Src_Zh	**25.5**(±1.5)	22.6(±1.5)	24.4(±1.0)	22.6(±1.4)
Src_En	22.9(±2.4)	23.1(±1.6)	**23.6**(±2.3)	23.1(±1.6)
Src_ZhEn	**26.5**(±4.3)	25.3(±3.1)	23.8(±1.0)	24.2(±1.2)
Src_EnZh	**25.8**(±1.9)	24.4(±0.9)	24.3(±1.1)	23.8(±0.8)

(b) the target task (CCKS) labels as query templates

Train\Test	Tgt_Zh	Tgt_En	Tgt_ZhEn	Tgt_EnZh
Tgt_Zh	**23.8**(±1.1)	18.2(±7.6)	18.9(±1.5)	21.8(±4.3)
Tgt_En	21.7(±4.7)	**24.9**(±1.0)	22.0(±6.5)	21.3(±2.8)
Tgt_ZhEn	25.2(±4.6)	25.3(±3.9)	25.6(±1.2)	**26.9**(±1.2)
Tgt_EnZh	22.8(±2.1)	24.0(±1.2)	**24.3**(±2.7)	23.0(±1.1)

As can be seen in Table 6:

- The highest performance was achieved when using a mixture of English and Chinese labels of the target task as the query template(F1 value of nearly 41). This indicates that using labels that are semantically similar to the target task entities can better induce cross-lingual prior knowledge in the model.
- Whether using source task labels or target task labels as query templates, when both training and prediction utilize the same queries, the F1 performance metric generally outperforms other scenarios. This suggests that employing identical query templates for both training and prediction is advantageous for the model's induction of prior knowledge.
- When training and prediction are conducted using mixed-language queries, regardless of the order of Chinese and English, the transfer performance generally surpasses other scenarios. This indicates that the position of labels within the template has a relatively minor impact on the induction of prior knowledge.

The differences between the scenarios presented in Table 7 and those in Table 6 are shown as follows:

- In Table 7(a), during training, using source task labels that include Chinese as query templates, and during testing, employing the "Src_Zh" query template containing only Chinese, achieved relatively better performance. This might be attributed to the fact that the target task's text is in Chinese, and the labels from the source task (I2B2) are relatively broad and general.
- As observed in Table 6, the absence of achieve the optimum values, when training and prediction use the same query templates in Table 6. It may be attributed to the fact that the entities annotated in the I2B2 dataset are more generic compared to the CCKS dataset. Furthermore, mixed language queries induce more information in the model, allowing models trained on the I2B2 corpus to recognize a broader range of

entities. This results in more generic false positives when predicting the CCKS dataset, thereby having an impact on the model's performance.

Comparison of performance for different entity types.
To explore performance differences between different entity types in different transfer directions, we selected the highest performance values in two transfer directions for analysis. Table 8 compares the performance of different entity types under mixed-language query templates, with the highest values among the three entity types indicated in bold.

Table 8. Comparison of performance for different entity types.

Entity Type	Tgt_ZhEn (CCKS to I2B2)			Tgt_ZhEn (I2B2 to CCKS)		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Medical problem	**75.7**	21.2	33.1	15.6	44.0	23.1
Medical treatment	72.9	**38.5**	**50.4**	29.2	34.6	31.7
Medical test	59.5	30.3	40.1	**34.8**	**45.0**	**39.3**
Micro Avg	68.3	29.1	40.8	19.3	44.3	26.9

As shown in Table 8:

- Although the "Medical problem" type has the largest proportion in both datasets, it has the lowest F1 score in both transfer directions, and it is lower than the overall F1 score. This is due to the semantic differences between the annotated entities in the two corpora, and the more entities there are, the greater the impact of noise on the transfer.
- The "Medical treatment" entity achieved the highest performance in the transfer direction from CCKS to I2B2, but it performed poorly in the reverse direction. This is because the I2B2 training set contains too many broad concept entities, which have a negative impact on the model's transfer effectiveness.
- The performance of "Medical test" did not vary significantly in both transfer directions, mainly due to the relatively small semantic differences in the annotation of "Medical test" entities between the two corpora. Additionally, "Medical test" entities appear in a relatively fixed format, and a considerable portion of entities in the Chinese dataset are represented using English abbreviations, such as "CT".

Performance comparison with baseline systems.
In Table 9, we compare our method with several commonly used methods in Cross-lingual NER.BDS_BERT(Bio_Discharge_Summary_BERT)[24] and Chinese_BERT_wwm[25] represent the best monolingual encoder models in Chinese and English, respectively. We employ cross-lingual word alignment information to project the source language into the target language and treat the task as monolingual NER. For a fair comparison, we also introduce the MRC framework into their methods. The

XLM-R model refers to the direct model transfer using sequence labeling on a cross-lingual pretrained model.

Our proposed method is divided into two categories: "Sgl", where query templates contain only one language, and "Mix", where query templates contain both languages. The performance in the table corresponds to the highest values for these two approaches. Similarly, the highest Precision/Recall/F1 scores among these methods are represented in bold.

Table 9. Performance comparison with baseline systems.

(a) CCKS to I2B2

Model	P(%)	R(%)	F1(%)
BDS_BERT+MRC	64.7	**29.7**	40.7(±0.8)
XLM-R	54.7	27.7	36.8(±2.2)
XLM-R+MRC(Sgl)	65.5	28.5	39.7(±1.7)
XLM-R+MRC(Mix)	**68.3**	29.1	**40.8(±1.6)**

(b) I2B2 to CCKS

Model	P(%)	R(%)	F1(%)
Chinese-BERT-wwm+MRC	15.2	**58.8**	24.1(±0.7)
XLM-R	13.2	56.8	21.4(±0.6)
XLM-R+MRC(Sgl)	17.7	45.4	25.5(±1.5)
XLM-R+MRC(Mix)	**19.3**	44.3	**26.9**(±1.2)

- After adopting the MRC framework, the model's transfer performance in both transfer directions significantly outperformed the sequence labeling approach, demonstrating the advantages of MRC in cross-lingual named entity recognition tasks.
- In both transfer directions, XLM-R+MRC with mixed language query templates achieved the highest F1 values among all baseline systems. Compared to using single-language templates, it obtained a positive improvement of 1.05 and 1.46, demonstrating the effectiveness of the mixed-query approach in cross-lingual pretrained models.
- Our proposed XLM-R+MRC(Mix) approach showed comparable performance to BDS_BERT+MRC and a significant improvement over the Chinese-BERT-wwm+MRC method. This is because BDS_BERT was pretrained on clinical domain text, endowing the model with domain-specific knowledge. When combined with MRC, it can better utilize prior knowledge to induce domain-specific knowledge into the model, thereby enhancing task performance.

6 Discussion and Case Study

6.1 Mixed language query and single language query

To investigate the reasons behind the improved model transfer performance of mixed language query templates, we selected the settings with the highest values achieved using mixed language queries and single-language queries in both transfer directions for comparison. The highest values in the comparison results are indicated in bold, as shown in Table 10.

Table 10. Comparison of single and mixed templates.

(a) CCKS to I2B2

Entity Type	Tgt_En			Tgt_ZhEn		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Medical problem	**76.3**	19.6	31.2	75.7	**21.2**	**33.1**
Medical treatment	70.5	36.8	48.3	**72.9**	**38.5**	**50.4**
Medical test	55.3	**32.3**	**40.8**	**59.5**	30.3	40.1
Micro Avg	65.5	28.5	39.7	**68.3**	**29.1**	**40.8**

(b) I2B2 to CCKS

Entity Type	Tgt_En			Tgt_ZhEn		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Medical problem	11.1	59.1	18.7	**15.6**	44.0	**23.1**
Medical treatment	**30.6**	**42.2**	**35.5**	29.2	34.6	31.7
Medical test	34.8	**48.0**	**40.4**	**34.8**	45.0	39.3
Micro Avg	16.4	**51.7**	24.9	**19.3**	44.3	**26.9**

From Table 10, we can observe the following:

- The use of mixed language query templates results in a more noticeable improvement in precision, particularly for the "Medical problem" and "Medical treatment" entity types. This suggests that mixed language query templates, compared to single-language templates, enable the model to acquire more prior knowledge to enhance the accuracy of predicting entities.
- In the CCKS to I2B2 direction, the results generally exhibit a "high precision, low recall" pattern, whereas in the I2B2 to CCKS direction, a "high recall, low precision" scenario is observed. This is due to the semantic differences in entities and concepts annotated in the two monolingual datasets. The broad concepts annotated in I2B2 lead to more false positives when transferred to CCKS, while the fine-grained entities annotated in CCKS result in the recognition of some fine-grained entities within the broad concepts when transferred to I2B2, leading to the opposite pattern.

6.2 Source task labels and target task labels

To investigate the reasons behind the improved model transfer performance using source task label templates, we selected the settings with the highest values achieved using source task labels and target task labels in both transfer directions for comparison. The highest values in the comparison results are indicated in bold, as shown in Table 11.

Table 11. Comparison of source and target task labels.

(a) CCKS to I2B2

Entity Type	Src_ZhEn			Tgt_ZhEn		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Medical problem	72.0	17.3	27.9	**75.7**	**21.2**	**33.1**
Medical treatment	71.6	36.3	48.2	**72.9**	**38.5**	**50.4**
Medical test	51.5	**32.9**	**40.2**	**59.5**	30.3	40.1
Micro Avg	63.0	27.6	38.7	**68.3**	**29.1**	**40.8**

(b) I2B2 to CCKS

Entity Type	Src_ZhEn			Tgt_ZhEn		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Medical problem	14.3	**52.5**	22.4	**15.6**	44.0	**23.1**
Medical treatment	23.8	**37.2**	29.1	**29.2**	34.6	**31.7**
Medical test	**35.4**	43.3	39.0	34.8	**45.0**	**39.3**
Micro Avg	**20.1**	39.0	26.5	19.3	**44.3**	**26.9**

In both transfer directions, using target task labels contributes to an improvement in recall and enhances transfer performance. Employing labels that are similar to the target corpus as queries aids the model in capturing the relationship between prior knowledge and context entities. For example, in the I2B2 dataset sentence, "She also received Cisplatin 35 per meter squared on 06/19 and Ifex and Mesna on 06/18", using "Src_ZhEn" did not identify "Ifex" and "Mesna" entities, while "Tgt_ZhEn" recognized all of them. The machine reading comprehension framework assists the model in capturing the relationship between the prior knowledge "Medical treatment" and the context word "received", thereby inducing the model to recognize more correct entities and enhancing transfer performance.

7 Conclusion

In this paper, we constructed a corpus for cross-lingual clinical named entity recognition (CLC-NER) using label alignment on existing monolingual datasets, demonstrating the effectiveness of the mixed-language query approach. Given that the semantic differences in annotated entities in the corpus limit the model's transfer performance, manual annotation of cross-lingual NER data in the clinical domain is necessary in future research.

Reference

1. Mayhew, S., Tsai, C.-T., Roth, D.: Cheap Translation for Cross-Lingual Named Entity Recognition. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2536–2545. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/D17-1269.
2. Xie, J., Yang, Z., Neubig, G., Smith, N.A., Carbonell, J.: Neural Cross-Lingual Named Entity Recognition with Minimal Resources, http://arxiv.org/abs/1808.09861, (2018).
3. Wu, S., Dredze, M.: Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. 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). pp. 833–844. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1077.
4. Tsai, C.-T., Mayhew, S., Roth, D.: Cross-Lingual Named Entity Recognition via Wikification. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. pp. 219–228. Association for Computational Linguistics, Berlin, Germany (2016). https://doi.org/10.18653/v1/K16-1022.
5. Keung, P., Lu, Y., Bhardwaj, V.: Adversarial Learning with Contextual Embeddings for Zero-resource Cross-lingual Classification and NER. 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). pp. 1355–1360. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1138.
6. Sang, T.K., Erik, F.: Introduction to the conll-2002 shared task: language-independent named entity recognition. In: Proceedings of CoNLL-2002/Roth, Dan [edit.]. pp. 155–158 (2002)
7. Sang, E.T.K., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. pp. 142–147 (2003)
8. Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., Ji, H.: Cross-lingual Name Tagging and Linking for 282 Languages. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1946–1958. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1178.
9. Ni, J., Dinu, G., Florian, R.: Weakly Supervised Cross-Lingual Named Entity Recognition via Effective Annotation and Representation Projection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1470–1480. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1135.
10. Jain, A., Paranjape, B., Lipton, Z.C.: Entity Projection via Machine Translation for Cross-Lingual NER. 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). pp. 1083–1092. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1100.
11. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, p. 2 (2019)

12. Wu, Q., Lin, Z., Karlsson, B., Lou, J.-G., Huang, B.: Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6505–6514. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.581.
13. Wu, Q., Lin, Z., Karlsson, B.F., Huang, B., Lou, J.-G.: UniTrans : Unifying Model Transfer and Data Transfer for Cross-Lingual Named Entity Recognition with Unlabeled Data. Presented at the Twenty-Ninth International Joint Conference on Artificial Intelligence July 9 (2020). https://doi.org/10.24963/ijcai.2020/543.
14. Li, Z., Hu, C., Guo, X., Chen, J., Qin, W., Zhang, R.: An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 170–179 (2022).
15. Levy, O., Seo, M., Choi, E., Zettlemoyer, L.: Zero-Shot Relation Extraction via Reading Comprehension. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). pp. 333–342. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/K17-1034.
16. Rajpurkar, P., Jia, R., Liang, P.: Know What You Don"t Know: Unanswerable Questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 784–789. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-2124.
17. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A HUMAN GENERATED MACHINE READING COMPREHENSION DATASET. (2017).
18. Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J.: A Unified MRC Framework for Named Entity Recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5849–5859. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.519.
19. Cao, J., Zhou, X., Xiong, W., Yang, M., Du, J., Yang, Y., Li, T., et al.: Electronic medical record entity recognition via machine reading comprehension and biaffine. Discrete Dynamics in Nature and Society 2021, 1–8 (2021)
20. Zheng, H., Qin, B., Xu, M.: Chinese Medical Named Entity Recognition using CRF-MT-Adapt and NER-MRC. Presented at the 2021 2nd International Conference on Computing and Data Science (CDS) January 1 (2021). https://doi.org/10.1109/CDS52072.2021.00068.
21. Han, X., Wang, Z., Zhang, J., Wen, Q., Li, W., Tang, B., Wang, Q., Feng, Z., Zhang, Y., Lu, Y., et al.: Overview of the ccks 2019 knowledge graph evaluation track: entity, relation, event and qa. arXiv preprint arXiv:2003.03875 (2020)
22. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 18, 552–556 (2011). https://doi.org/10.1136/amiajnl-2011-000203.
23. CONNEAU, A., Lample, G.: Cross-lingual language model pretraining. Advances in Neural Information Processing Systems 32 (2019)
24. Cao, J., Zhou, X., Xiong, W., Yang, M., Du, J., Yang, Y., Li, T., et al.: Electronic medical record entity recognition via machine reading comprehension and biaffine. Discrete Dynamics in Nature and Society 2021, 1–8 (2021)
25. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-Training With Whole Word Masking for Chinese BERT. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 3504–3514 (2021). https://doi.org/10.1109/TASLP.2021.3124365.