
北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

# 命名实体抽取

组员： 徐亚苗， 华玲誉， 张博， 林宁， 李雪宜

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

# 目录

CONTENTS


1 基本介绍

2 主要方法

3 应用实例

## 命名实体（named entity，NE）：

1995 年 11 月的第六届 MUC 会议 （MUC-6， the Sixth Message Understanding Conferences）上被提出的。

**MUC**：规定了 NER 评测需要识别的三大类（命名实体、时间表达式、数量表达式）、七小类实体。其中命名实体分为：人名、机构名和地名。

**ACE**：将命名实体中的机构名和地名进行了细分，增加了地理-政治实体和设施两种实体，之后又增加了交通工具和武器。

**CoNLL**：将命名实体定义为包含名称的短语，包括人名、地名、机构名、时间和数量，实际 的任务主要是识别人名、地名、机构名。

## 命名实体识别（named entity recognition，NER）：

识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。

| 小明 | 在 | 北京 | 看了一场 | 中国男篮 | 的比赛。 |
| --- | --- | --- | --- | --- | --- |
| PER | | LOC | | ORG | |

命名实体识别是实体关系抽取、信息检索、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分。

**过程：** （1）实体边界识别；（2） 确定实体类别（人名、地名、机构名或其他）。

**难点：**

(1)命名实体类型多样,数量众多,不断有新的命名实体涌现，难以建立大而全的数据库。

(2)命名实体构成结构比较复杂，存在大量的嵌套、别名、缩略词等问题,没有严格的规律可以遵循;人名中也存在比较长的少数民族人名或翻译过来的外国人名。

(3)不同命名实体之间界限不清晰,人名也经常出现在地名和组织名称中,存在大量的交叉和互相包含现象。常常要涉及上下文语义层面的分析,这些都给命名实体的识别带来困难。

(4)在不同的文化、领域、背景下,命名实体的外延有差异。对命名实体的定界和类型确定,目前还没有形成共同遵循的严格的命名规范。

(5)命名实体识别过程常常要与中文分词、浅层语法分析等过程相结合,分词、语法分析系统的可靠性也直接决定命名实体识别的有效性,使得中文命名实体识别更加困难。

## 命名实体识别工具：

| | | |
|---|---|---|
| 国外 | Stanford NER | 斯坦福大学开发的基于条件随机场的命名实体识别系统，该系统参数是基于CoNLL、MUC-6、MUC-7和ACE命名实体语料训练出来的。 |
| | NLTK | 由宾夕法尼亚大学开发的自然语言处理工具包，在NLP领域中最常使用的一个Python库。 |
| | MALLET | 麻省大学开发的一个统计自然语言处理的开源包，其序列标注工具的应用中能够实现命名实体识别。 |
| 国内 | Hanlp | 作者何晗，HanLP是一系列模型与算法组成的NLP工具包，目标是普及自然语言处理在生产环境中的应用。支持命名实体识别。 |
| | LPT | 哈尔滨工业大学开发的中文自然语言处理工具。提供了一整套自底向上的丰富而且高效的中文语言处理模块。 |
| | ICTCLAS | 中国科学院计算技术研究所，当前世界上最好的汉语词法分析器，支持命名实体识别。 |

# 早期方法：基于规则和词典



人工构建有限的规则，再从文本中寻找匹配这些规则的字符串

**01**

虽然能够在特定的语料上获得较高的识别效果。但是识别效果越好，越需要大量规则的制定，而人工制定这些规则可行性太低。

**02**

几乎不可能通过制定有限的规则来识别出变化无穷的命名实体。

**03**

规则对领域知识极度依赖，当领域差别很大时，制定的规则往往无法移植，不得不重新制定规则。

这些固有的缺点使得研究者们转而采取新的研究思路，而此时正值机器学习在 NLP 领域兴起，NER也自然地转向了机器学习的阵营。

# 2 传统方法：基于机器学习



基于机器学习的NER的方法归根结底是分类的方法。给定命名实体的多个类别，再使用模型对文本中的实体进行分类。分为两种思路：

1

先识别出文本中所有命名实体的边界，再对这些命名实体进行分类。

2

序列化标注方法。利用大规模语料学习出标注模型，再对句子的各个位置进行标注。常用的模型有隐马尔可夫模型(Hidden Markov Model，HMM)和条件随机场(Conditional Random Field，CRF)等。

# 传统方法：基于机器学习



隐马尔可夫模型（Hidden Markov Model，HMM）创建于上世纪70年代，是一个统计模型。HMM的系统中存在两条序列：

- 一个是可以直接通过观测得到的观察序列，在NER中指每一个词语本身；
- 另一个是隐含的状态转移序列，指每个词语背后的标注。

我们要做的就是求观察序列的背后最可能的标注序列。即根据输入的一系列单词，去生成其背后的标注，从而得到实体。

# 传统方法：基于机器学习



条件随机场（ConditionalRandom Field，CRF）是NER目前的主流模型。假设$X$表示待标记的观测序列，$Y$表示隐状态序列，$P(Y|X)$表示给定$X$的条件下$Y$的条件概率，随机变量$Y$满足：

$$P(Y_v|X,Y_w,w\neq v)=P(Y_v|X,Y_w,w\sim v)$$ 对任意的$v$都成立

节点$v$对应的随机变量

节点$v$以外的所有的节点$w$

节点$w$对应的随机变量

节点$v$的所有邻接节点$w$

即，对于一个结点$v$,它与其他所有与它不邻接的节点相互独立。

# 传统方法：基于机器学习

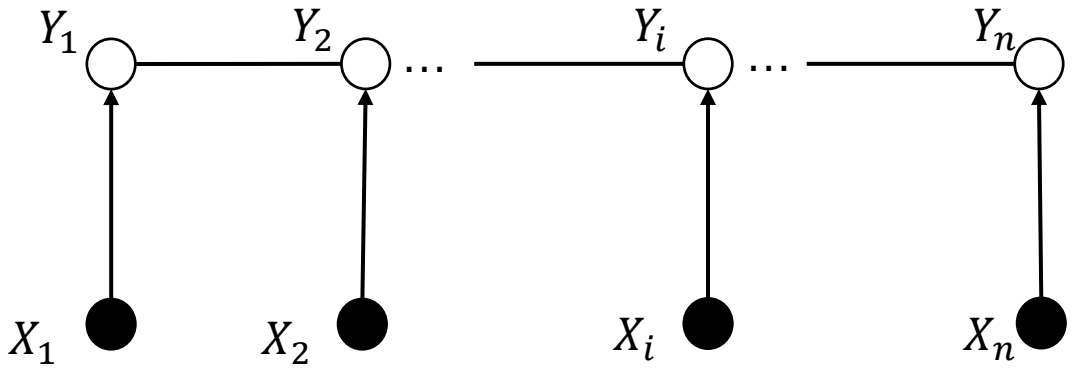


在NER任务中，使用的主要是链式结构的CRF，即线性链条件随机场（linear chain CRF）

假设输入的观测序列$X=(X_1, X_2, \ldots, X_n)$，对应的状态序列$Y=(Y_1, Y_2, \ldots, Y_n)$，在给定随机变量序列$X$的条件下，随机变量序列$Y$的条件概率分布$P(Y|X)$构成条件随机场，满足：

$$P(Y_i|X, Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n) = P(Y_i|X, Y_{i-1}, Y_{i+1}).$$

$i = 1,2, \ldots, n$ (在i=1和n时只考虑单边)



# 2 近期方法：基于神经网络



近年来，随着硬件计算能力的发展以及word embedding的提出，神经网络已经成为一个可以有效处理许多NLP任务的模型。在NER任务中，传统的神经网络的处理方式：

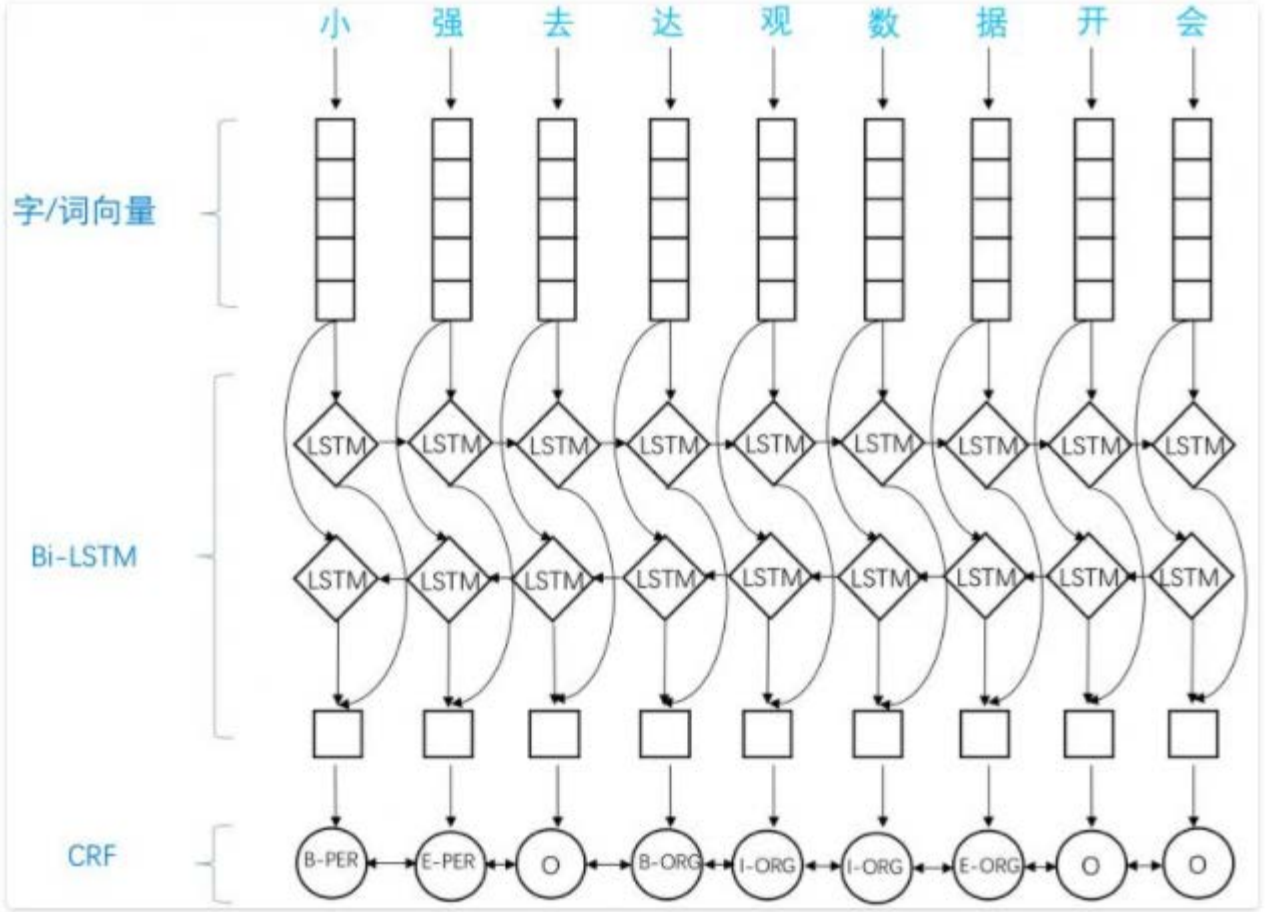
- 先将句子中的所有单词表示为word embedding；
- 随后将句子的embedding序列输入到神经网络中，用神经网络自动提取特征；
- 最后通过一层Softmax来预测每个token的标签。

其中，神经网络通常采用RNN，LSTM，BiLSTM等。

# 近期方法：基于神经网络



然而，这种传统的神经网络方法对每个单词打标签的过程是一个独立的分类，不能直接利用上文已经预测的标签，这可能导致预测出的标签序列是非法的。为解决这一问题，学界提出了“神经网络-CRF”模型做序列标注。比较有代表性的是BiLSTM-CRF：

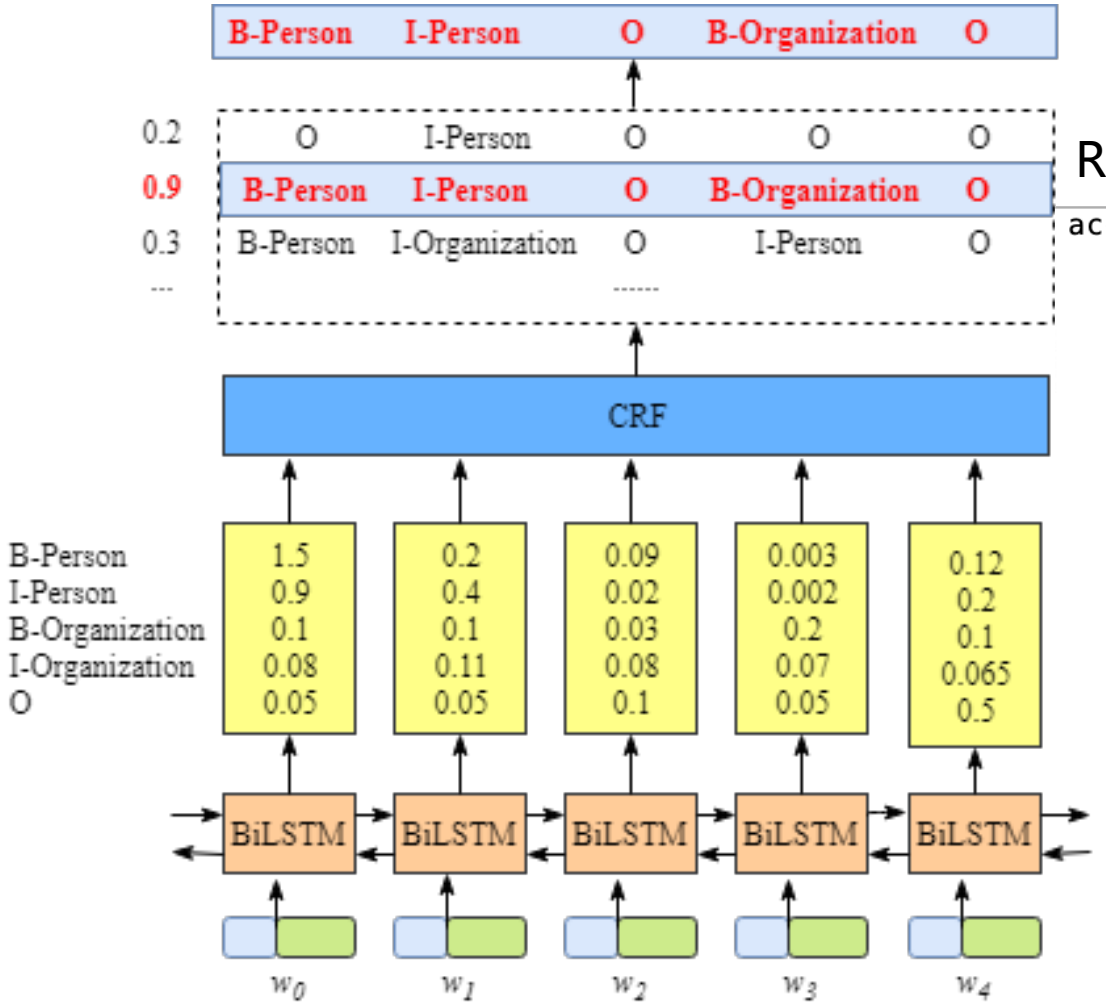


- 在CoNLL 2003数据集上使用三种现有的NER模型进行训练与评测(BiLSTM+CRF,BiLSTM+CNN+CRF,BERT+CRF);

- 在中文NER数据集上使用BERT+CRF模型进行训练与评测;

# BiLSTM-CRF



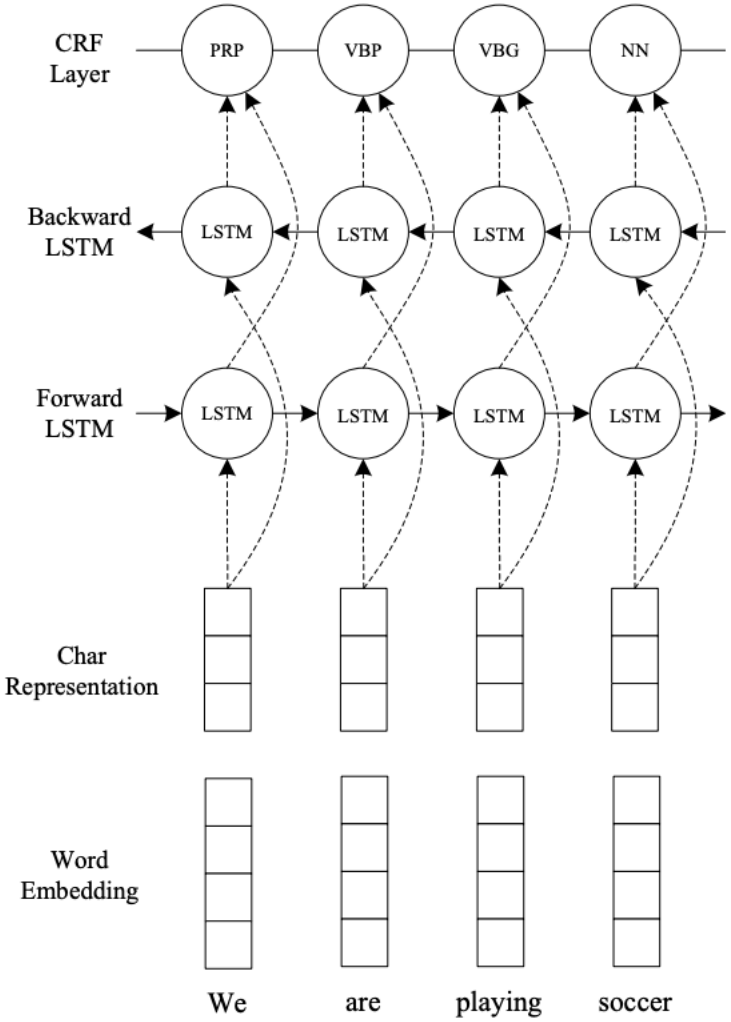


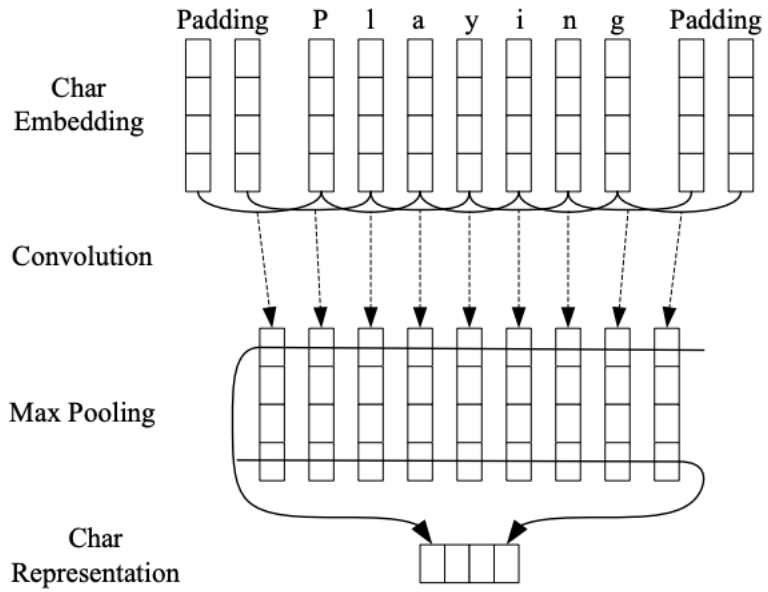

RESULTS:(On test set)

```
accuracy:  98.15%; precision:  90.61%; recall:  88.85%; FB1:  89.72
              LOC: precision:  91.93%; recall:  91.79%; FB1:  91.86  1400
             MISC: precision:  83.83%; recall:  78.43%; FB1:  81.04  713
              ORG: precision:  87.83%; recall:  85.18%; FB1:  86.48  1253
              PER: precision:  95.19%; recall:  94.83%; FB1:  95.01  1294
```

# BiLSTM-CNN-CRF





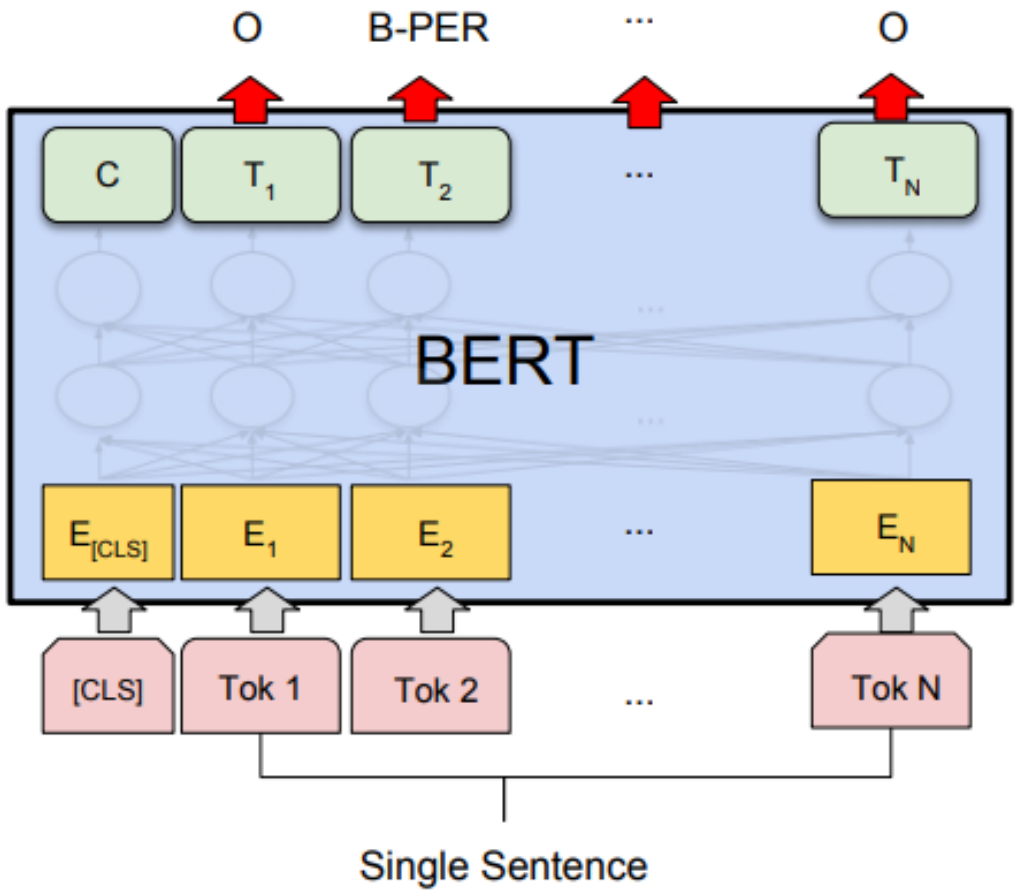




RESULTS:(On test set)

```
accuracy:  98.16%; precision:  91.33%; recall:  89.97%; FB1:  90.12
              LOC: precision:  92.15%; recall:  91.79%; FB1:  91.92  1400
             MISC: precision:  83.97%; recall:  80.93%; FB1:  82.01  713
              ORG: precision:  87.89%; recall:  86.72%; FB1:  87.35  1253
              PER: precision:  95.36%; recall:  94.76%; FB1:  95.76  1294
```

# BERT-CRF






RESULTS:(On test set)

```
accuracy:  98.83%; precision:  91.36%; recall:  91.02%; FB1:  91.19
              LOC: precision:  92.56%; recall:  91.90%; FB1:  92.63  1400
             MISC: precision:  86.67%; recall:  85.79%; FB1:  86.23  713
              ORG: precision:  91.36%; recall:  89.72%; FB1:  90.34  1253
              PER: precision:  95.12%; recall:  92.95%; FB1:  93.36  1294
```

# 模型比较



CoNLL2003以F1为评测标准，三种模型的F1如下表：

| Model | % P | % R | % F1 |
|---|---|---|---|
| BLSTM-CRF | 90.61 | 88.85 | 89.72 |
| BLSTM-CNN-CRF | 91.33 | 89.97 | 90.12 |
| BERT-CRF | 91.36 | 91.02 | 91.19 |

对比三种模型的F1值可知，在CoNLL 2003数据集上BERT-CRF的NER效果更好。

# 3 BERT-CRF在中文NER数据集上的表现



BERT也提供了中文的预训练模型：

- `BERT-Large, Uncased (Whole Word Masking)`: 24-layer, 1024-hidden, 16-heads, 340M parameters
- `BERT-Large, Cased (Whole Word Masking)`: 24-layer, 1024-hidden, 16-heads, 340M parameters
- `BERT-Base, Uncased`: 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Large, Uncased`: 24-layer, 1024-hidden, 16-heads, 340M parameters
- `BERT-Base, Cased`: 12-layer, 768-hidden, 12-heads , 110M parameters
- `BERT-Large, Cased`: 24-layer, 1024-hidden, 16-heads, 340M parameters
- `BERT-Base, Multilingual Cased (New, recommended)`: 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Base, Multilingual Uncased (Orig, not recommended)` **(Not recommended, use `Multilingual Cased` instead)**: 102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Base, Chinese`: Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters ✓

我们基于BERT中文预训练模型在中文NER上进行训练与测试：

RESULTS:(On test set)

```
accuracy:  99.37%; precision:  95.74%; recall:  95.33%; FB1:  95.53
              LOC: precision:  95.91%; recall:  96.05%; FB1:  95.98  3469
              ORG: precision:  93.05%; recall:  92.15%; FB1:  92.60  2145
              PER: precision:  98.61%; recall:  97.75%; FB1:  98.18  1804
```

# BERT-CRF在中文NER数据集上的表现



所得模型的NER效果：

```
input the test sentence:
据新华社报道，日前，国家卫健委、国家发改委等12部门联合发布《关于加快落实仿制药供应保障及使用政策工作方案》提出，
2019年6月底前，发布第一批鼓励仿制的药品目录。
[['O','B-ORG','I-ORG','I-ORG','O','O','O','O','O','O','B-ORG','I-ORG','I-ORG','I-ORG','I-
ORG','O','B-ORG','I-ORG','I-ORG','I-ORG','I-ORG','O','O','O','O','O','O','O','O','O','O',
'O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O
','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O',
'O','O']]
LOC
PER
ORG, 新华社, 国家卫健委, 国家发改委
input the test sentence:
而且冯绍峰爆料在拍摄《知否》的那段时间里，赵丽颖就吃很少，很瘦，也根本没有胃口吃，一整天钻在剧本中无法自拔。
[['O','O','B-PER','I-PER','I-PER','O','O','O','O','O','O','O','O','O','O','O','O','O','O',
'O','O','B-PER','I-PER','I-PER','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O',
'O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O','O
','O','O','O']]
LOC
PER, 冯绍峰, 赵丽颖
ORG
input the test sentence:
国务院总理李克强在雄安新区召开会议
[['B-ORG','I-ORG','I-ORG','O','O','B-PER','I-PER','I-PER','O','B-LOC','I-LOC','I-LOC','I-
LOC','O','O','O','O']]
LOC,雄安新区
PER,李克强
ORG,国务院
```