# 基于深度学习的命名实体识别综述

## 命名实体识别任务

### 1.1 命名实体识别概念

在文本中对实体的引用可以有三种形式：命名性指称、名称性指称、代词性指称。例如在句子“**中国乒乓球男队主教练刘国梁出席了会议，他布置了备战世乒赛的具体安排**”中“刘国梁”的指称项有三个，其中“中国乒乓球男队主教练”是名词性指称，“刘国梁”是命名性指称，“他”是代词性指称。

命名实体识别任务是识别出文本中实体的命名性指称项，并标明其类别。一般来说，命名实体识别任务中，实体类别一般分为三大类（实体类、时间类、数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）。不同任务对命名实体类别粒度的需求不同，例如在有些任务中只要识别出一个实体是人，有些任务则需要识别出一个实体是教师、学生或者医生。

### 命名实体识别标注方法

目前主流的命名实体标注方法分为两种，一种是BIO方法,一种是BIOES方法。BIO方法中B标注命名实体的开始词、I标注命名实体识别的中间词、O标注不属于命名实体的其他词。BIOES方法中B、I、O的意义与BIO标注方法中的相同，但增加了E和S，E标注命名实体的结束词、S标注只包含一个词的命名实体。通常在标注的过程中，会加上命名实体的类别信息。如下是分别使用BIO方法和BIOES方法对同一个句子进行标注的结果。

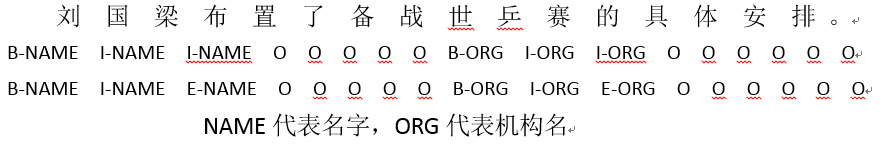


图1.1

### 命名实体识别评估标准

在命名实体识别任务中，只有当预测的命名实体边界和类别与实际完全相同时，才可以认为识别成功。准确率、召回率，F1值是评估命名实体识别系统性能的三大指标。式（1-1）、式（1-2）、式（1-3）分别是它们的计算公式。其中TP(True Positive)表示预测的实体中，正确实体的数量、FP(False Positive)表示预测的实体中错误实体的数量。FN(False Negative)表示语料库中未被预测出的实体数量。

 式（1-1）

 式（1-2）

 式（1-3）

## 基于深度学习的命名实体识别方法

传统的命名实体识别方法有基于规则的命名实体识别方法、基于机器学习的命名实体识别的方法。以上两种都需要依靠大量的人力去构建模板或者特征，而基于深度学习的方法 ，依靠深度学习强大的非线性表达能力，不需要人力去构建模板或者特征，可以实现端到端的模式【】，成为当前命名实体识别方法的主流。

基于深度学习的命名实体识别方法，都遵循着通用的框架，如图2.1所示。通用框架可以分为三个串行的模块，分别是输入的分布式表示、内容编码、标签解码。输入的分布式表示指的是单词、字符的嵌入还有词性、外部知识库的分布式表示。内容编码是通过使用CNN、RNN或者的其他的神经网络获得内容之间的依赖信息。标签解码是预测对应词的标签。

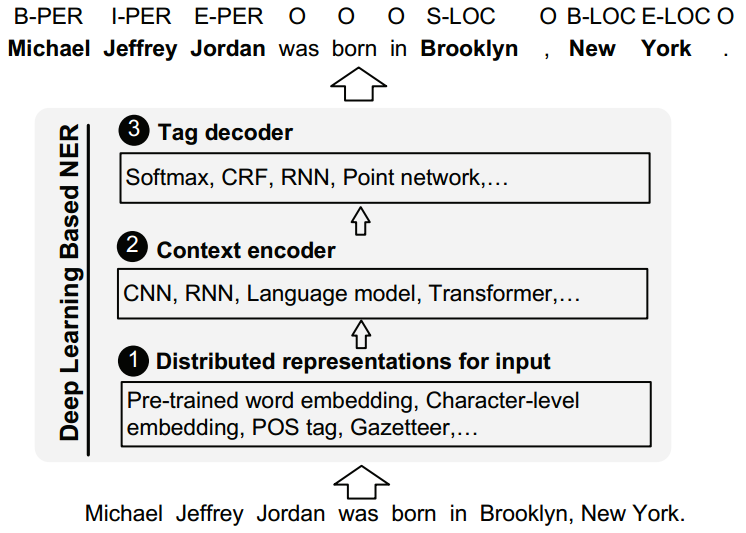


图2.1

### 2.1输入的分布式表示

基于深度学习的命名实体识别方法的输入一般是低维稠密的实值向量，所以需要获得输入的分布式表示。输入的分布式表示一般包括词嵌入、字嵌入，同时一些方法也会考虑词的词性信息、。

#### 2.1.1词嵌入

词嵌入主要分为两类，分别是基于词向量的固定表征包括word2vec、fastText、glove，基于词向量的动态表征包括elmo、GPT、Bert。

word2vec是Mikolov于2013年提出的词嵌入方法，包括skip-gram模型和CBOW模型。

skip-gram模型的原理是给定输入词来预测上下文。CBOW模型则相反是给定上下文来预测输入词。图2.2是两种模型的网络结构。

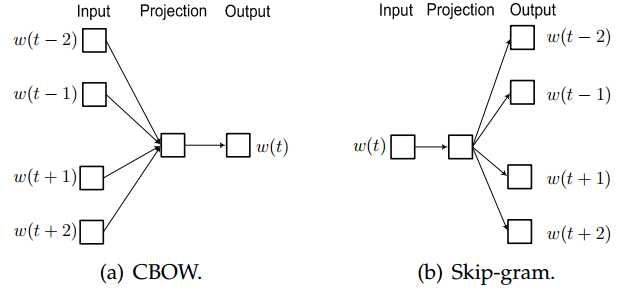


图2.2

Glove【】是Jeffrey与2014年提出的词嵌入方法，是一种半统计半机器学习的方法。与word2vec方法仅考虑输入词附近的窗口信息不同，Glove通过构建词与词之间的窗口共现矩阵考虑了全局的信息。

FastText【】是Mikolov于2016年提出的，在skip-gram基础上考虑词的内部结构信息。

FastText不仅对词本身编码也对其字符级的n-gram特征进行编码，语料中的低频词，可能与某些高频词具有相同的n-gram特征，所以可以得到可靠的向量表示。

Elmo是基于词向量的动态表征方法。上面提到的word2vec、glove、fasttext都是基于词向量的静态表征方法，每一个词都只有一个唯一的向量与之对应。多义词在不同语境中有不同的意思，难以用一个词向量对其所有意思进行编码。为了克服这个缺点，Elmo分为两个阶段对词进行编码，第一阶段用CNN提取词字符级特征，作为第二阶段的原始输入 ，第二阶段利用两层双向LSTM提取上下文的语境信息，最后整合原始输入和两层的双向LSTM的输出，构建最后的词向量。在Elmo方法中，词向量不是固定的，而是根据具体的上下文语境信息，动态的确定。

GPT、Bert同样也是基于词向量的动态表征方法。与Elmo不同的是，它们都利用Transformer而不是LSTM来提取上下文依赖信息(实践表明, Transformer的特征提取能力要强于LSTM)。GPT与Bert之间的差异性主要体现在GPT采用的是单向语言模型，而Bert采用的是双向语言模型。

#### 2.1.2 字符嵌入

相关研究证明，在英文中，考虑词的内部结构信息（前缀、后缀等），可以使命名实体识别系统的性能提升。现在主流一般用基于CNN的模型或者基于RNN的模型提取词的字符级特征。如图2.3所示。是两种模型的网络结构。Lample提出的基于BiLSTM+CRF的命名实体识别方法是使用双向LSTM做字符嵌入。Chiu提出的基于LSTM-CNNs的命名实体识别方法是使用CNN做字符嵌入。同样Elmo方法中，也是用CNN提取词的字符级特征。

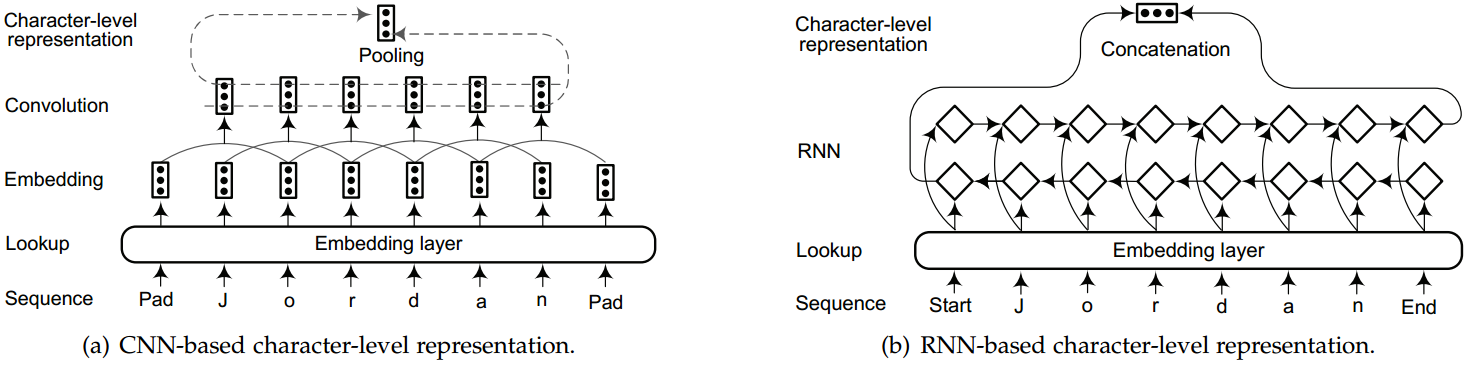


图2.3

### 内容编码

### 2.3标签解码