**中文命名实体识别：**

**任务描述**

首先，我们明确一下命名实体识别的概念：

命名实体识别（英语：Named Entity Recognition），简称NER，是指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，以及时间、数量、货币、比例数值等文字。

举个例子，假如有这么一句话：





**为什么要识别命名实体**

命名实体是现实世界里的事物，它们和现实世界相互作用、相互影响，因此命名实体在一些场景里特别重要，比如我想知道《红楼梦》里所有的人物在各回的出场情况，进而为每个人物做一个生平简介。这时候，我们就需要某种手段，把文本中的命名实体给识别出来。

有时候，我们需要知道事物之间的关系，进而准确地决策。比如说，王夫人需要知道大观园里所有人和宝玉的交往情况，找出所有可能引诱宝玉不走正道的“狐媚子”。我们可以识别出大观园里出现的所有的人物，并判断人物之间的关系(有想法、没想法)，然后做成一个有向图。这样，就可以把所有对宝玉有想法的人拎出来审问了。

这么做似乎代价有点高啊。如果只是为了保护宝玉，做这个图谱就有点铺张了；如果是为了保护所有的人，这个图谱的成本就会被摊的比较薄，值得做。

**如何识别命名实体**

**1.人工标注**

命名实体是人定义的，人当然可以胜任这个工作。但是，就像前面所说的，有几个限制因素，导致不能依靠人工来做NER：（1）做事情需要人、财、物，而人力资源是其中最金贵的，耗费比较大；（2）我们在标注数据的时候，会面临体力下降、情绪波动等等生理和心理状况的考验，导致不能长时间、高质量地进行数据标注；（3）我们处理数据的速度太慢了，这是最要命的。举个例子，我标注微博文本的情感极性时，一天上千条就烦的不行——生产力太弱。

如果数据量比较小，使用真正的“人工智能”是可以的；当数据量比较大的时候，我们需要机器的帮助。我们一般用人工标注一个足够大的高质量训练数据，然后训练模型来做大规模的NER。

**2.使用规则**

规则是最符合我们直觉的一种算法，任何需要自动化的地方，我们都会尝试用规则来解决问题，NER也不例外。在人工智能发展的初期，人们认为语言规律是可以用规则来描述的，比如“xxx公司”是一个公司名。

一般来说，我们在做命名实体的时候，可以首先考虑一下可否使用正则。假如命名实体的名称规律比较简单，我们可以找出模式，然后设计相应的正则表达式或者规则，然后把符合模式的字符串匹配出来，作为命名实体识别的结果。比如我需要识别文本里的政府机构。这里的机构名称规律比较简单，是“[地名][职能][行政级别]”这样的三元组(然后去掉地名，就是我们要的部门名称了)。符合这个模式的，“基本上”就是一个政府机构。为什么说是“基本上”，因为我只看了很少的数据，总结的模式还很少。

**3.基于词典**

词典是一种非常有价值的数据。在命名实体识别任务中，我们可以把部门名称当做一个模式，直接去文本里匹配——如果一个部门名称出现在文本里，说明文本包含了这个名称。

这样做会遇到比较严重的问题：歧义。比如“我国的自然资源局部集中现象很普遍”，这句话里的“自然资源局”是一个部门名称吗？不是。

我们可以用分词的方法，来减少歧义带来的困扰。先将句子切分为“我/国/的/自然/资源/局部/集中/现象/很/普遍”,然后逐词匹配即可。当然，这要求分词算法比较给力。

那要是遇到“我是自然资源局的”，分词结果是“我/是/自然/资源/局/的”,这可怎么办呢？“自然资源局”被切分成若干个词语了。这时候可以使用最大匹配法，检查分词结果中是否在这样的一些子串，可以组成部门名称。

实际上，现在比较流行的分词工具(HanLP,Jieba,IK等)，默认用的都是最短路径分词这类算法，支持用户添加自定义词典、使特定字符串优先成词。这样，我们把部门名称添加的词典里，“自然资源局”就不会被切分开了。

**4.模型**

后来，由于无法解决语言相关的问题，基于规则的人工智能系统退出了C位。取而代之的是统计和机器学习。

**1）分类器**

在一些人的眼中，万物皆可分类。NER也可以看做一种分类任务，就是判断一个词语是不是命名实体、是哪一种命名实体。常见的做法就是，基于一个词语或者字的上下文构造特征，来判断这个词语或者字是否为命名实体。

这个方法需要比较好的特征工程，也就是要求我们对文本和业务内容有很深的理解，门槛还是比较高的。

**2）HMM和CRF**

NER也可以用序列标注的方式来做。命名实体存在于自然语言中，而自然语言是“人”这个模型在接收外界信号后生成的一种序列，因此有人认为语言符合某种模型。

常见的一种假设是，序列元素具有一种隐藏（不可见）的状态——模型以一种概率分布随机生成隐藏状态，然后基于隐藏状态的取值选择一种概率分布取生成序列。在NER中，词语的标签就是不可见（因此需要推测）的状态。我们可以罗列出所有可能生文字序列或者词语序列的NER标签序列，然后用HMM和CRF评估各个标签序列的质量、择优录取（实际不需要这么暴力，人们为这两个模型提供了缩小搜索空间的算法）。

这两个模型的优点是，使用了文本序列的整体信息，可以找到“最合适的”标注方案。因此，即使深度学习来势汹汹，CRF也没有被淘汰。

**3）深度学习**

深度学习领域的模型结构种类比较多，最适合做自然语言处理任务的是RNN类和tansformer类。这些神经网络也把语言看做是序列数据，然后用自身极强的拟合能力，把这种序列转换为标签序列。

如果只使用神经网络，我们会用softmax来作为输出层，本质上是对序列的元素进行分类。这种做法比较自然，不过有一个不足：它认为序列元素之间的相互独立的，损失了不少信息。

还有一种做法，就是以CRF作为输出层，把任务变成序列标注，这样就可以使用上序列元素之间的关系信息了。这个方案结合了神经网络的拟合能力和CRF的全局视野，是非常经典、有效的一种NER模型结构。

当然了，去年开始大家开始关注的transformer做NER更是一把好手。我们可以把transformer和CRF结合起来。是的，特征提取器总是在变，而CRF一直在顶端。

**涉及模型：**

BiLSTM + CRF

BiLSTM + CNN + CRF

Transformer

**论文复现：**

* Chinese NER Using Lattice LSTM ACL 2018
* Simplify the Usage of Lexicon in Chinese NER ACL 2020
* FLAT: Chinese NER Using Flat-Lattice Transformer ACL 2020
* Named Entity Recognition for Social Media Texts with Semantic Augmentation EMNLP 2020

**数据集整理：**

MSRA NER:

数据概览：5万多条中文命名实体识别标注数据（包括地点、机构、人物）

下载地址：<https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/dh_msra/intro.ipynb>

Resume:

数据集情况：由新浪金融高管简历组成

下载地址：<https://pan.baidu.com/s/1p8YgbZgXdedHzfNpn5ntrg>

提取码：hell

Weibo:

数据集情况：由微博社交媒体文本组成

<https://pan.baidu.com/s/1QfWukFze_modzz3vo9cWBw>

提取码：hell

OntoNotes 4.0:

you can download it in [LDC](https://www.ldc.upenn.edu/)