**实体识别**

1. **人民日报语料2014版 - pd2014**

2014年更新更全的一版人民日报语料。分成两个数据文件，前一个是分词后的原文本，后一个是分词后一一对应的标签，总共大概28.6万个句子。样本如下：

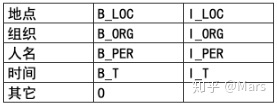
==> source\_BIO\_2014\_cropus.txt <==

人 民 网 1 月 1 日 讯 据 《 纽 约 时 报 》 报 道 , 美 国 华 尔 街 股 市 在 2 0 1 3 年 的 最 后 一 天 继 续 上 涨 , 和 全 球 股 市 一 样 , 都 以 最 高 纪 录 或 接 近 最 高 纪 录 结 束 本 年 的 交 易 。

==> target\_BIO\_2014\_cropus.txt <==

O O O B\_T I\_T I\_T I\_T O O O B\_LOC I\_LOC O O O O O O B\_LOC I\_LOC I\_LOC I\_LOC I\_LOC O O O B\_T I\_T I\_T I\_T I\_T 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

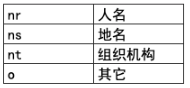
这份数据采用的标签集（BIO标记法）如下：



<https://github.com/hspuppy/hugbert/blob/master/ner_dataset/pd2014.tar.bz2>

1. **MSRA数据集 - msra**

由微软亚洲研究院提供的一份数据集，训练集加测试集大概5万句左右。样例如下： 4 我们/o 藏/o 有/o 一/o 册/o １９４５年６月/o 油印/o 的/o 《/o 北京/ns 文物/o 保存/o 保管/o 状态/o 之/o 调查/o 报告/o 》/o ，/o 调查/o 范围/o 涉及/o 故宫/ns 、/o 历博/ns 、/o 古研所/nt 、/o 北大清华图书馆/ns 、/o 北图/ns 、/o 日/ns 伪/o 资料/o 库/o 等/o 二十几家/o ，/o 言/o 及/o 文物/o 二十万件/o 以上/o ，/o 洋洋/o 三万/o 余/o 言/o ， /o 是/o 珍贵/o 的/o 北京/ns 史料/o 。/o 注意到它的分词不是字粒度的。语料中用到的标注如下，需要适当的脚本转化成为BIO标记法。



<https://github.com/hspuppy/hugbert/blob/master/ner_dataset/msra.tgz>

1. **微博数据集 - weibo**

由微博提供，并且已经转成了多种（基本）直接可用的格式，数据量大概1500句。

其中coll格式（后面用HF的transformers库也要转成类似的格式）的数据样本如下： 欧 B-GPE.NAM 美 I-GPE.NAM 时 O 尚 O 热 O 门 O 榜 O 心 O 谢 B-PER.NAM 耳 I-PER.NAM 朵 I-PER.NAM 。 O 这部分数据也可以从这里下载：<https://github.com/hspuppy/hugbert/blob/master/ner_dataset/weibo.tgz>

微博这份数据使用了较多的标签：



由于标签多，这项任务也相对较难。

1. **BOSON数据集 - boson**

由玻森中文语义开放平台提供，共2000句。(https://bosonnlp.com/ 已经停止服务)

样例： 此次{{location:中国}}个展，{{person\_name:苏珊?菲利普斯}}将与她80多岁高龄的父亲一起合作，哼唱一首古老的{{location:威尔士}}民歌{{product\_name:《白蜡林》}}。届时在{{location:画廊大厅}}中将安放6个音箱进行播放，艺术家还特意回到家乡{{location:格拉斯哥}}，同父亲一起在{{org\_name:中国音乐学院}}里为作品录制了具有{{location:中国}}元素的音乐片段。 它采用了双花括号的方式标注，实体类别共6种：**time: 时间 location: 地点 person\_name: 人名 org\_name: 组织名 company\_name: 公司名 product\_name: 产品名。**

<https://github.com/hspuppy/hugbert/blob/master/ner_dataset/boson.tar.gz>

1. **CLUE数据集 - clue**

CLUE中提供了一份细粒度NER数据集，“CLUE Fine-Grain NER”，训练集10748，验证集1343 。格式为json： {"text": "浙商银行企业信贷部叶老桂博士则从另一个角度对五道门槛进行了解读。叶老桂认为，对目前国内商业银行而言，", "label": {"name": {"叶老桂": [[9, 11]]}, "company": {"浙商银行": [[0, 3]]}}} {"text": "生生不息CSOL生化狂潮让你填弹狂扫", "label": {"game": {"CSOL": [[4, 7]]}}} 它共提供了10类实体标签，在网站上有详细说明，分别是：

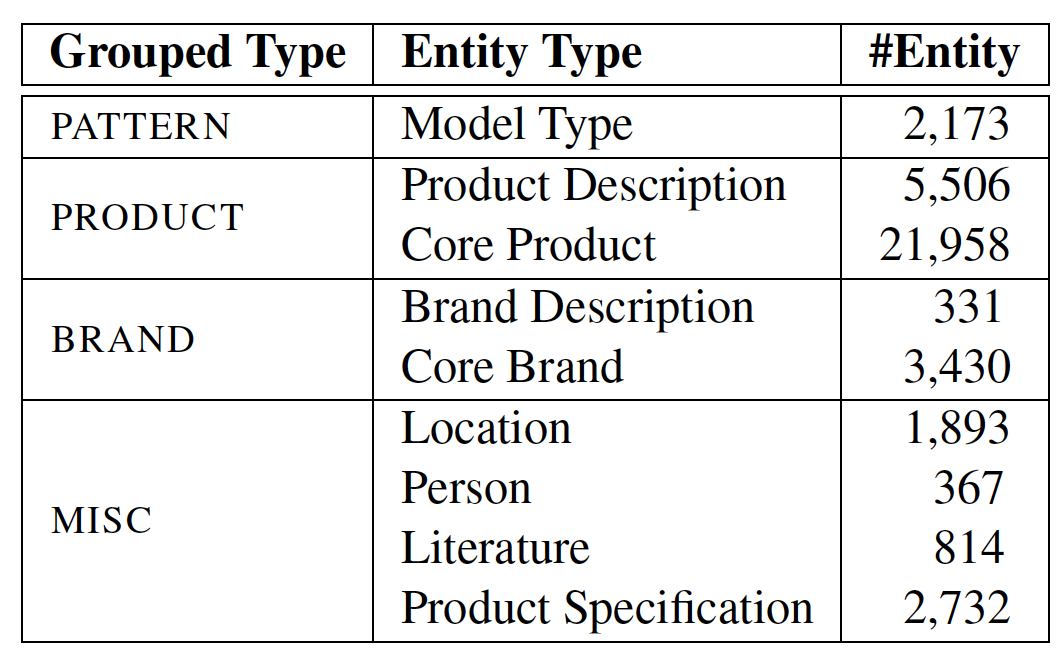
**地址（address），书名（book），公司（company），游戏（game），政府（goverment），电影（movie），姓名（name），组织机构（organization），职位（position），景点（scene**） 。

<https://github.com/hspuppy/hugbert/blob/master/ner_dataset/clue.tgz>

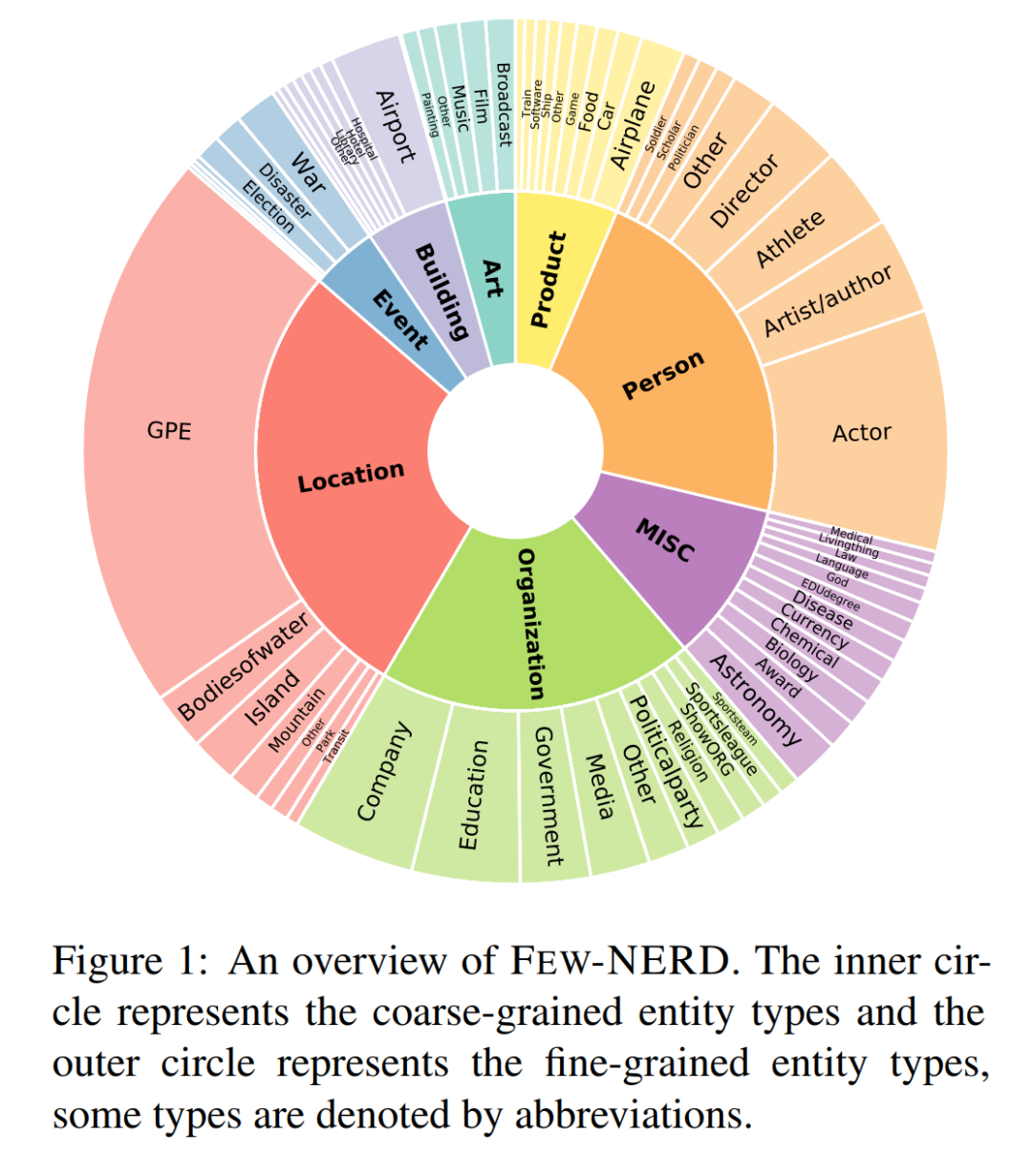
1. E-Commercial NER Dataset / 电商NER数据集

命名体识别(NER)是一项重要的自然语言处理任务，本数据集提供了电商领域的NER开放数据集，包括了4大类、9小类实体类别。该数据集由阿里巴巴达摩院和新加坡科技设计大学联合提供。

https://tianchi.aliyun.com/dataset/dataDetail?dataId=108758



1. Few-NERD：一个Few-shot场景的命名实体识别数据集



Few-NERD，一个大规模的人工标注的用于few-shot NER任务的数据集。该数据集包含8种粗粒度和**66种细粒度实体**类型，每个实体标签均为粗粒度+细粒度的层级结构https://www.jiqizhixin.com/articles/2021-06-03。

**关系抽取**

https://www.yuque.com/docs/share/fa05692a-7340-4bb6-9a14-5d1357a2b

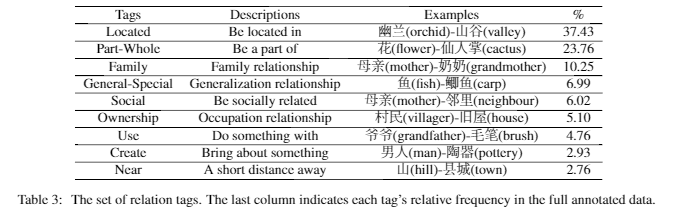
38f

1. 中文实体关系抽取数据集

基于几个可用的NER和RE数据集定义了7个实体标签和9个关系标签，但还有一些针对汉语文学的附加类别。

<https://www.heywhale.com/mw/dataset/5dde487dca27f8002c4a8352>

[https://github.com/lancopku/Chinese-Literature-NER-RE-Dataset](https://github.com/lancopku/Chinese-Literature-NER-RE-Dataset" \t "https://www.heywhale.com/mw/dataset/_blank)



**关系抽取**是从纯文本中识别实体之间的关系事实，也是完成实体识别任务后的一个重要工作，从抽取的范围来看，包括句子级RE关系和文档级关系。同样的，当前也陆续开放出来了一些实体关系评测和数据集，例如ACE、semeval、Duee等。

下面整理得到10大类开放数据集。

1、ACE实体关系数据集

ACE2005数据集包括英语，阿拉伯语和中文三部分数据，分成广播新闻和新闻专线两部分。

该数据集包含ART、Gen-affiliation在内的6种关系类型，Employment、Founder、Ownership在内的额18种子关系类型。

ACE04语料库共包含451个文档和5 702个关系实例。ACE2005中文数据集包括633个文档、307991个字符。

地址： https://catalog.ldc.upenn.edu/byproject

2、SemEval实体关系数据集

SemEval数据集是2010年国际语义评测大会中Task8任务所使用的数据集，共包括Cause-Effect(因果关系)、Instrument-Agency(操作、使用关系)、Product-Producer(产品-生产者关系)、 Content-Container(空间包含关系)、Entity-Origin(起源关系)、Entity-Destination(导向关系)、 Component-Whole(组件-整体关系)、Member-Collection(成员-集合关系)、Message-Topic(主题关系)等10类关系。

该数据集包括8000个训练样本，2717个测试样本。

地址： https://github.com/thunlp/OpenNRE/blob/master/benchmark/download\_semeval.sh

3、FewRel实体关系数据集

FewRel是清华大学于2018年发布的精标注关系抽取数据集，是当前规模最大的中文实体关系数据集。

该数据集包括100个关系类别、70,000个关系实例。每句的平均长度为24.99，一共出现 124,577 个不同的单词/符号。

地址： https://github.com/thunlp/OpenNRE/blob/master/benchmark/download\_fewrel.sh

4、Wiki80实体关系数据集

Wiki80是从数据集FewRel上提取的一个关系数据集，共包含location、part of、follows等80种关系，每种关系个数均为700，共56000个样本。

该数据集50400条训练集、5600条验证集。

地址： https://github.com/thunlp/OpenNRE/blob/master/benchmark/download\_wiki80.sh

5、NYT10实体关系数据集

NYT10是在基于远程监督的关系抽取任务上最常用的数据集，由NYT corpus 在2010年基于Freebase远程监督得到的，共包含founders、place\_of\_birth在内的53种关系（包括一种NA），存在一定的噪声。

该数据集包括466876条训练集、55167条验证集以及172448条测试集。

地址： https://github.com/thunlp/OpenNRE/blob/master/benchmark/download\_nyt10.sh

6、DuIE2.0实体关系数据集

DuIE2.0是业界规模最大的基于schema的中文关系抽取数据集，来自百度百科、百度贴吧和百度信息流文本。

该数据集包含超过43万三元组数据、21万中文句子及48个预定义的关系类型。

该数据集包括171135个训练集、21055个测试数据，外加80184条混淆数据。

地址： https://github.com/PaddlePaddle/PaddleNLP/tree/develop/examples/information\_extraction/DuIE

7、COAE2016实体关系数据集

CAOE2016关系数据集是CAOE2016 task3任务中用到的一个关系数据集，涉及到任关系类别包括出生日期、出生地、毕业院校、配偶、子女、高管、员工数、创始人、总部、其他共十类关系。

该数据集中包含988个训练数据和483个测试数据。 地址： https://ccir2016.ccnl.scut.edu.cn/caoe\_test.php

8、人物实体关系数据集

人物关系抽取数据集是CCKS2019中的一个层级关系分类任务，包括三大类(亲属关系、社交关系、师生关系)，四中类(配偶、血亲、姻亲、友谊）、35小类(现夫、前妻)种关系类型。

该数据集包括3841条验证集、287351条训练集以及77092条测试集句子。

地址： https://github.com/SUDA-HLT/IPRE

9、Chinese Literature Text文档级实体关系数据集

Chinese Literature Text是面向中文文学的一个实体关系数据集，标注了物体、人名、地名、时间名、容量名、组织和摘要共7类实体，位于、部分、家庭、概括、社会、拥有、使用、制造、邻接等9类实体关系。

该数据集共计726篇文章，29096句话，超过100000个字符。训练集695篇，验证集58篇、测试集84篇。

地址： https://github.com/lancopku/Chinese-Literature-NER-RE-Dataset

10、DocRED文档级实体关系数据集

DocRED是基于维基百科的文档级关系抽取数据集，数据集中每个文档都被标注了命名实体提及、核心参考信息、句内和句间关系以及支持证据。关系类型涉及科学、艺术、时间、个人生活在内的96种Wikidata关系类型。

该数据集在5053个维基百科文档上进行标注，包含132375个实体和56354个关系事实。

地址： https://github.com/thunlp/DocRED