实体识别任务简介

命名实体识别概念

命名实体识别(Named Entity Recognition,简称NER)，是指识别文本中具有特定意义的词（实体），主要包括人名、地名、机构名、专有名词等等,并把我们需要识别的词在文本序列中标注出来。

例如有一段文本:李明在天津市空港经济区的税务局工作

我们要在上面文本中识别一些区域和地点,那么我们需要识别出来内容有:

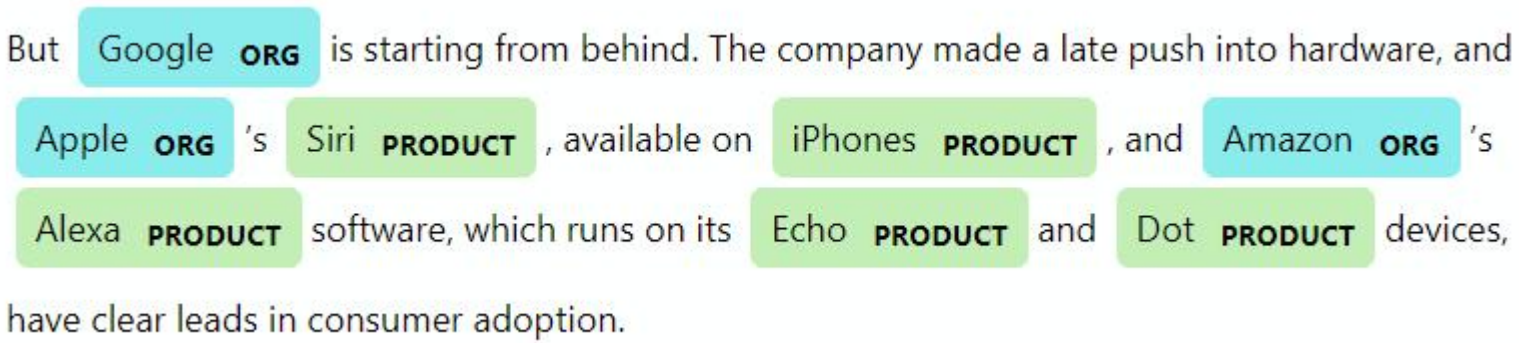
李明(人名)、天津市(地点)、 空港经济区(地点)、税务局(组织)

识别上述例子我们使用了以下几个标签：

- "B-ORG":组织或公司(organization)
- "I-ORG":组织或公司
- "B-PER":人名(person)
- "I-PER":人名.
- "O":其他非实体(other)
- "B-LOC":地名(location)
- "I-LOC":地名

命名实体识别标签

NER的识别靠的是标签，在长期使用过程中，有一些大家使用比较频繁的标签，下面给出大家一些参考：



命名实体识别标注

在序列标注中，我们想对一个序列的每一个元素(token)标注一个标签。一般来说，一个序列指的是一个句子，而一个元素(token)指的是句子中的一个词语或者一个字。比如信息提取问题可以认为是一个序列标注问题，如提取出会议时间、地点等。

标签类型的定义一般如下：

定义	全称	备注
B	Begin	实体片段的开始
I	Intermediate	实体片段的中间
E	End	实体片段的结束
S	Single	单个字的实体
O	Other/Outside	其他不属于任何实体的字符(包括标点等)

命名实体识别标注

BIO标注模式

将每个元素标注为“B-X”、“I-X”或者“O”。其中，“B-X”表示此元素所在的片段属于X类型并且此元素在此片段的开头，“I-X”表示此元素所在的片段属于X类型并且此元素在此片段的中间位置，“O”表示不属于任何类型。

命名实体识别中每个token对应的标签集合如下:

LabelSet = {O, B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG}

命名实体识别标注

BIOES标注模式

BIOES标注模式就是在BIO的基础上增加了单字符实体和字符实体的结束标识, 即

LabelSet = {O, B-PER, I-PER, E-PER, S-PER, B-LOC, I-LOC, E-LOC, S-LOC, B-ORG, I-ORG, E-ORG, S-ORG}

命名实体识别标签

NER的识别靠的是**标签**，在长期使用过程中，有一些大家使用比较频繁的标签，下面给出大家一些参考：

```
PERSON:      People, including fictional.
NORP:        Nationalities or religious or political groups.
FAC:         Buildings, airports, highways, bridges, etc.
ORG:         Companies, agencies, institutions, etc.
GPE:         Countries, cities, states.
LOC:         Non-GPE locations, mountain ranges, bodies of water.
PRODUCT:     Objects, vehicles, foods, etc. (Not services.)
EVENT:       Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART: Titles of books, songs, etc.
LAW:         Named documents made into laws.
LANGUAGE:    Any named language.
DATE:        Absolute or relative dates or periods.
TIME:        Times smaller than a day.
PERCENT:     Percentage, including ”%“.
MONEY:       Monetary values, including unit.
QUANTITY:    Measurements, as of weight or distance.
ORDINAL:     “first”, “second”, etc.
CARDINAL:    Numerals that do not fall under another type.
```

命名实体识别标签

NER的识别靠的是**标签**，在长期使用过程中，有一些大家使用比较频繁的标签，下面给出大家一些参考：

地点	B_LOC	I_LOC
组织	B_ORG	I_ORG
人名	B_PER	I_PER
时间	B_T	I_T
其它	O	

nr	人名
ns	地名
nt	组织机构
o	其它

地区名特指，如深圳	B-GPE.NAM I-GPE.NAM
地名特指，如华克山庄	B-LOC.NAM I-LOC.NAM
地名泛指，如寺庙	B-LOC.NOM I-LOC.NOM
组织名特指	B-ORG.NAM I-ORG.NAM
组织名泛指	B-ORG.NOM I-ORG.NOM
人名特指，如方进玉	B-PER.NAM I-PER.NAM
人名泛指，如男人	B-PER.NOM I-PER.NOM
其它	O

Grouped Type	Entity Type	#Entity
PATTERN	Model Type	2,173
PRODUCT	Product Description	5,506
	Core Product	21,958
BRAND	Brand Description	331
	Core Brand	3,430
MISC	Location	1,893
	Person	367
	Literature	814
	Product Specification	2,732

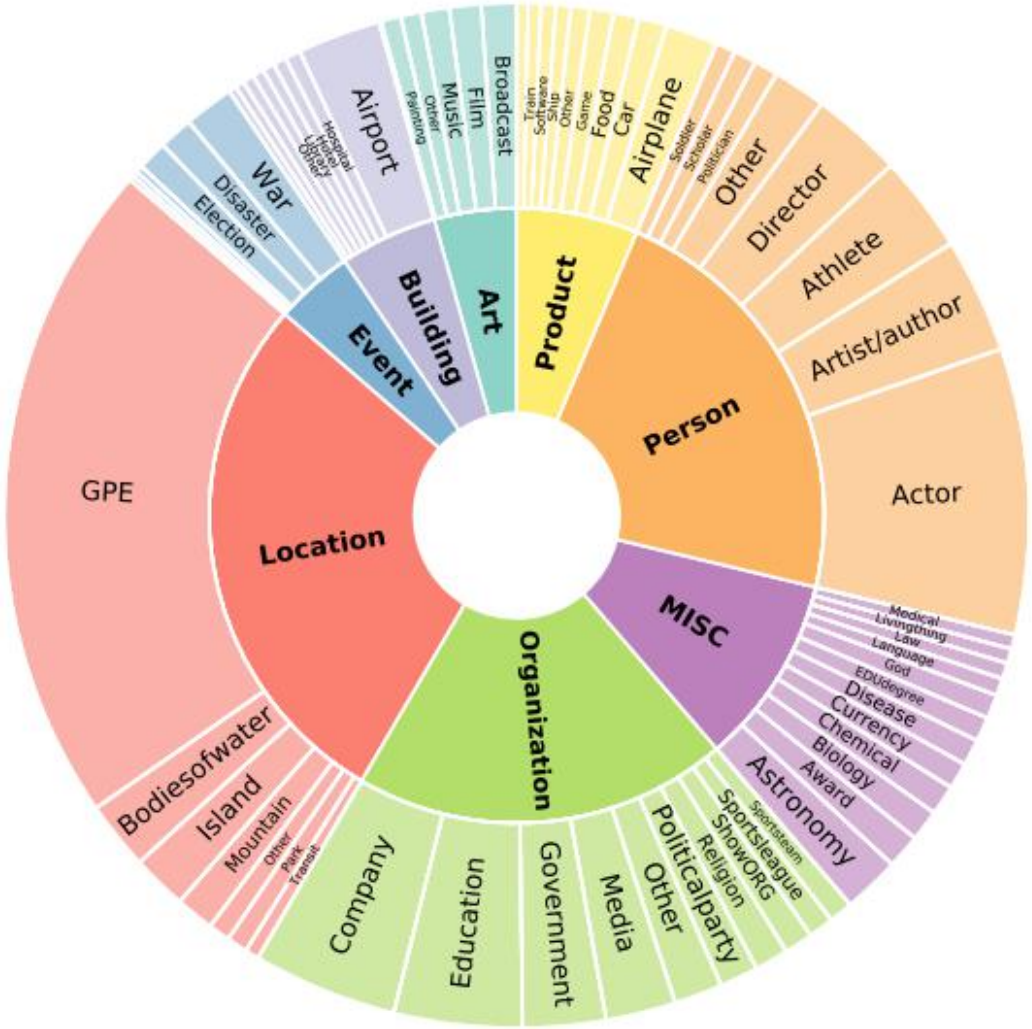
命名实体识别标签

NER的识别靠的是**标签**，在长期使用过程中，有一些大家使用比较频繁的标签，下面给出大家一些参考：

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。

FEW-NERD: A Few-shot Named Entity Recognition Dataset

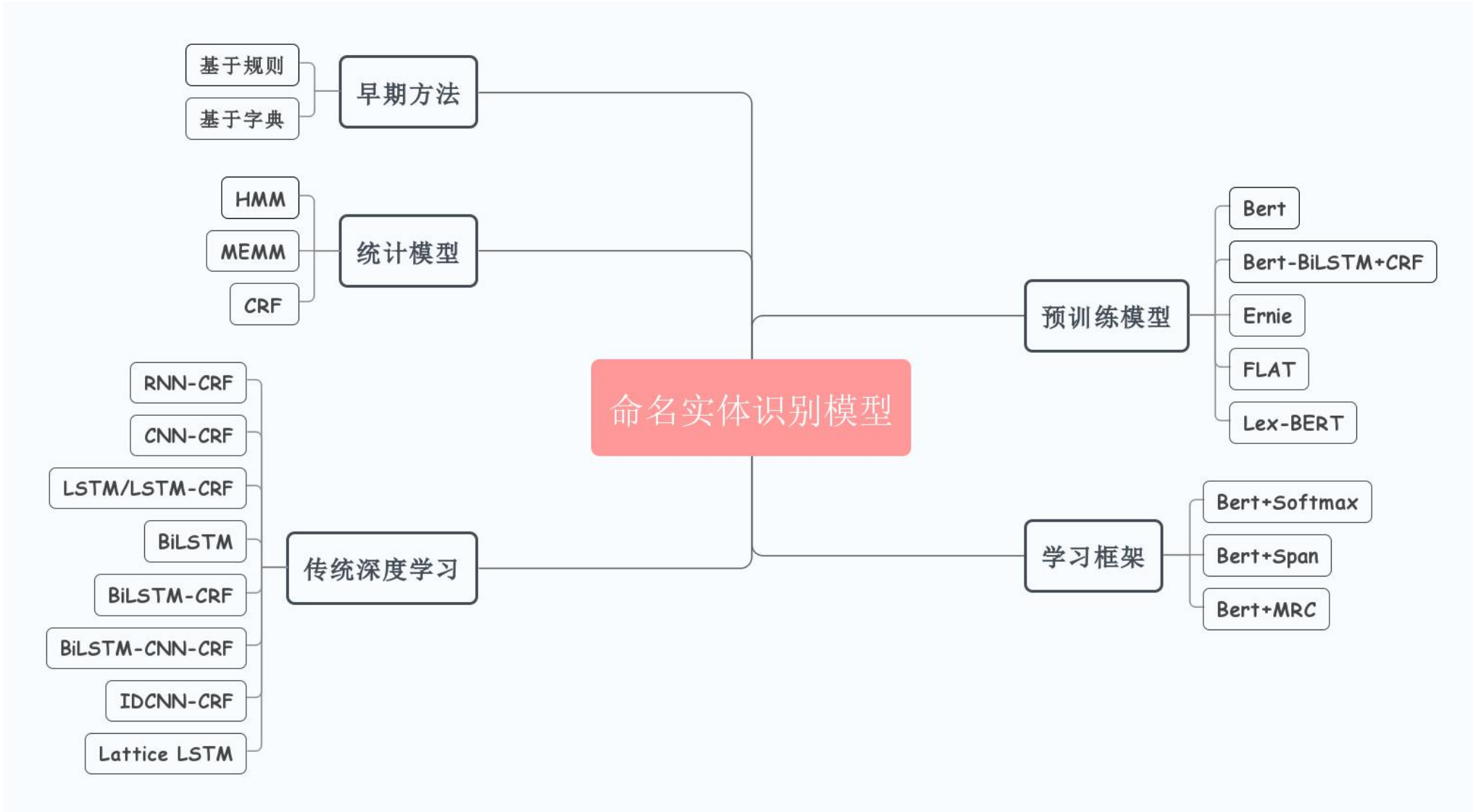
https://arxiv.org/abs/2105.07464



命名实体识别数据集

1、CLUENER2020：https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/cluener_public
2、MSRA：https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/MSRA
3、人民网（04年）：https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/people_daily
4、微博命名实体识别数据集：https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/weibo
5、BosonNLP NER数据：https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/boson（2000条）
6、影视-音乐-书籍实体标注数据：https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/video_music_book_
7、中文医学文本命名实体识别 2020CCKS：https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/2020_ccks_
8、电子简历实体识别数据集:https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/ResumeNER
9 、医渡云实体识别数据集:https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/yidu-s4k
10、 简历实体数据集：https://github.com/jiesutd/LatticeLSTM/tree/master/data
11、CoNLL-2003：https://www.clips.uantwerpen.be/conll2003/ner/
12、Few-NERD 细粒度数据集:https://github.com/thunlp/Few-NERD/tree/main/data
.....

命名实体识别模型


命名实体识别模型
早期方法
基于规则
基于字典
统计模型
HMM
MEMM
CRF
传统深度学习
RNN-CRF
CNN-CRF
LSTM/LSTM-CRF
BiLSTM
BiLSTM-CRF
BiLSTM-CNN-CRF
IDCNN-CRF
Lattice LSTM
预训练模型
Bert
Bert-BiLSTM+CRF
Ernie
FLAT
Lex-BERT
学习框架
Bert+Softmax
Bert+Span
Bert+MRC

命名实体识别工具

- 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

  https://nlp.stanford.edu/software/CRF-NER.shtml
  python实现的Github地址：https://github.com/Lynten/stanford-corenlp

- 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。
  官方地址：http://mallet.cs.umass.edu/

- Hanlp：HanLP是一系列模型与算法组成的NLP工具包，由大快搜索主导并完全开源，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。
  Github地址：https://github.com/hankcs/pyhanlp
  官网：http://hanlp.linrunsoft.com/

命名实体识别工具

- NLTK：NLTK是一个高效的Python构建的平台,用来处理人类自然语言数据。提供实体识别接口。
  Github地址：https://github.com/nltk/nltk
  官网：http://www.nltk.org/

- spaCy：工业级的自然语言处理工具。
  Gihub地址：https://github.com/explosion/spaCy
  官网：https://spcay.io/

- Crfsuite：可以载入自己的数据集去训练实体识别模型。
  文档地址：https://sklearn-crfsuite.readthedocs.io/en/latest/?badge=latest

- CRF++是基于C++开发、可自定义特征集、基于LBFGS快速训练等等高效特征的CRF开源工具包。用于对序列数据进行分割和标记，主要用于NLP任务，例如命名实体识别、信息提取和序列标注等任务。
  https://taku910.github.io/crfpp/

推荐资料

流水的NLP铁打的NER：命名实体识别实践与探索 - 知乎
https://zhuanlan.zhihu.com/p/166496466

中文NER的正确打开方式: 词汇增强方法总结 (从Lattice LSTM到FLAT)
https://zhuanlan.zhihu.com/p/142615620

自然语言处理基础技术之命名实体识别简介
https://www.jianshu.com/p/02b08ff8ad3c