南京農業大學
NANJING AGRICULTURAL UNIVERSITY

数字人文下的命名实体识别

目录
CONTENT

- 命名实体识别概念与基本原理
- 命名实体识别在数字人文中的应用
- 古文实体识别流程

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

命名实体识别
概念与基本原理

1、命名实体

◆实体识别所指的实体，是文本中的实体，是对广义实体的指称（Mention）。

◆指称的方式一般有三种

- **命名性指称（Name Mention）：通过名字来指称实体。**
- 名词性指称（Nominal Mention）：通过名词或名词性短语来表示实体。
- 代词性指称（Pronoun Mentions）：通过代词来指代实体。

誠 樸 篤力 仁

1、命名实体

◆命名实体是对实体的命名性指称，是一种专指性词项。

◆命名实体具有五种特性

- **指称性**：用来指示或称说某些事物，以便将这些事物跟其他事物区分开来。不是所有的词语都有指称性，例如形容词表示事物的性质，动词表示动作或行为。代词、名词通常都有指称性。
- **专门性**：专门用来指示或称说某一个事物，以便将这个事物跟同类的其他事物区分开来。例如，“教授”、“年轻的教授”都是对一类人的指称，而“李教授”则是对某一个姓李的教授的指称。（注意，“李教授”绝不是对所有姓李的教授的指称）。

誠樸篤仁

1、命名实体

- **词汇性：**命名实体属于词汇，词汇成员包括词和固定词组。组织名通常是固定词组，固定词组中一般不含虚词。凭句法手段构造的自由词组也可用来指称某个个体，例如，“这粒沙子”。这些自由词组不属于词汇，当然也不是命名实体。
- **开放性：**命名实体是词汇中最直接反映客观世界变化的部分。新事物不断产生，而且往往对我们特别有重要性，需要命名，所以命名实体的数量往往非常庞大，而且层出不穷，难以胜数。
- **可替换性：**每一类（或每一小类）中的命名实体之间是可以替换的。替换之后语法上、语义上仍然是成立的，尽管可能不符合事实。

誠樸篤仁

1、命名实体

◆人名

- 中国人名：马云、刘强东、张近东
- 外国人名译名：阿兰·图灵、冯·诺依曼

◆地名

- 中国地名：北京、上海、广州
- 外国地名译名：纽约、伦敦、巴黎

◆机构名

- 南京农业大学

1、命名实体

◆古汉语命名实体

- 古汉语命名实体研究中也常关注官职、谥号等，如官职“令尹”，谥号“桓”。古汉语命名实体研究较之现代汉语最大的差别在于，不仅关注以名字指称实体对象的“命名性指称”，还关注“名词性指称”，如“齐桓公”、“中行寅”、“聲孟子”等，这类人名中往往包含了谥号、官职等非名字成分，因此古汉语命名实体识别最大的难点之一就在于人名识别。

2、命名实体识别

- ◆命名实体识别（Named Entity Recognition，NER）是指通过设计相应的算法以序列和分类的思路实现对命名实体的识别。
- ◆其目的是识别出文本中表示命名实体的成分，并对其进行分类，因此有时也称为命名实体识别和分类（Named Entity Recognition and Classification，NERC）
- ◆命名实体识别任务早期源于信息抽取和信息检索的需求，因为命名实体经常成为检索关键词，并且是事件和关系中的重要结构项。后来逐渐成为自然语言处理中一项独立的重要任务。
- ◆未登录词识别是命名实体识别中的难点，而做好命名实体识别，也有助于提高未登录词识别的准确率和召回率。

誠樸勤仁

3、序列化标注

◆自然语言处理中常使用序列化标注的方法来完成命名实体识别任务。

◆在序列化标注中，首先将文本表示成词语或汉字的序列，然后使用机器学习模型对该序列中的每个词语或汉字进行分类。

◆序列化标注的类别有BIOES、BIO等模式。在BIOES类别模式下，B表示命名实体的开头，I表示命名实体内部，O表示命名实体之外的文本，E表示命名实体的结尾，S表示单独的命名实体。通过这样的类别模式完成了词语或汉字序列成分的分类之后，就相当于完成了命名实体识别的任务。

誠樸勤仁

4、特征

◆CRF、最大熵等机器学习模型完成序列化标注任务时，可以根据需要构建标注任务相关的特征函数，以提高标注任务的性能。

◆在命名实体识别任务中，特征函数的构建主要依据命名实体相关的语言知识，一般我们将这类知识叫做特征。

◆命名实体识别利用的特征可以分为全局特征和局部特征。

誠樸勤仁

4、特征

◆全局特征

- 命名实体（例如人名）常常有连续出现的情况，如果其中某个已经被识别为命名实体，利用搭配约束可提高识别其余命名实体的效果。
- 一个命名实体往往在初次出现时具有较丰富的上下文特征，以后出现时则不一定总带着这些特征。利用篇章约束可以提高其后续出现的识别效果。

4、特征

◆局部特征

- **构成特征**（命名实体内部）：在某种基元（词或字符）的序列中，命名实体为一子序列，充当这个子序列的各个基元及其属性。
- **上下文特征**（命名实体之外）：在某种基元（词或字符）的序列中，命名实体为一子序列，这个子序列之前或之后的基元及其属性。
- **语序特征**：带有位置信息的特征。例如当前词的前一个词（-1）的属性，后一个词（+1）的属性。
- **结构特征**：这里“结构”不限于语法结构，可宽泛理解为几个特征的同现或复合。例如前两个词（-2和-1）的复合属性，后两个词（+1和+2）的复合属性。

誠樸篤力仁

4、特征

◆探索、标注和测试各类特征知识，以及与此密切相关的特征工程，是机器学习时代命名实体识别中最重要的一项研究内容。

◆在深度学习兴起之后，知识的表示被词嵌入（word embedding）和预训练（pre-train）等表示学习方法替代，计算机可以自动从上下文中学到稠密低维的文本向量，削弱了特征知识标注的必要性，同时进一步提高了命名实体识别的效果。

◆深度学习模型与传统机器学习模型看待特征知识的视角截然不同，但这不意味着特征知识的探索的停止，有研究表明，利用好的特征知识的CRF模型能够获取比深度学习模型更好的性能，这在古汉语等语料规模较小的研究领域中尤其如此。

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

命名实体识别
在数字人文中的应用

命名实体识别在数字人文中的应用

◆命名实体识别是自然语言处理工作的基础性工作。面向现代文本的命名实体识别已经取得了丰硕的成果，于此同时，还有海量的数字化典籍资源有待深度挖掘。

◆当前命名实体识别工作在数字人文领域的应用主要集中于对文本的组织与利用和古籍实体识别系统的开发两方面。

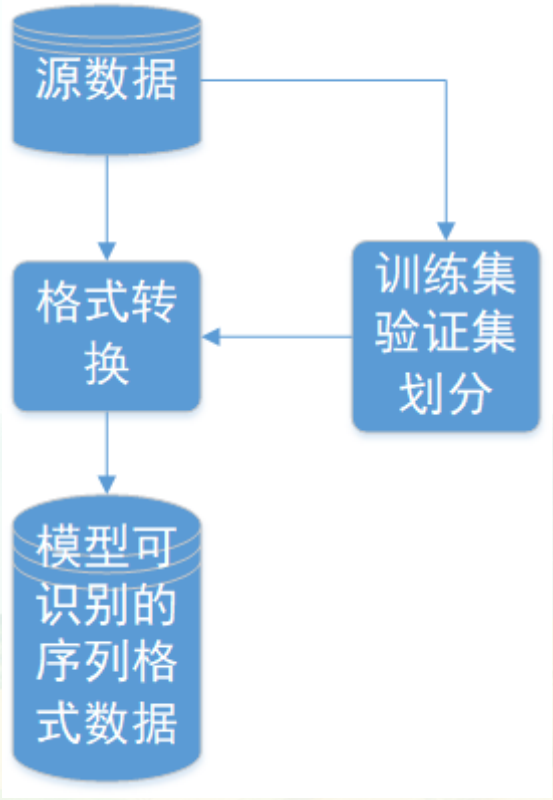
南京農業大學
NANJING AGRICULTURAL UNIVERSITY

古文实体识别流程

古文实体识别流程

◆1 数据的预处理

➢ 模型将实体识别问题看作一个序列标注问题，因而需要将标注好的典籍文本转换为模型能够识别的序列格式。



誠 樸 篤力 仁

古文实体识别流程

◆格式转换

输入格式

惠公/nr 元妃/v 孟子/nr 。/w ↓
孟子/nr 卒/v ， /w 繼/v 室/n 以/p 聲子/n ， /w 生/v 隱公/nr 。/w ↓
宋武公/nr 生/v 仲子/nr ， /w 仲子/nr 生/v 而/c 有/v 文/n 在/p 其/r 手/n ， /w 曰/v 為/v 魯/ns 夫人/n ， /w 故/c 仲子/nr 歸/v 于/p 我/r 。/w ↓

输出格式

惠 B-nr↓
公 E-nr↓
元 O↓
妃 O↓
孟 B-nr↓
子 E-nr↓
。 O↓
↓
孟 B-nr↓
子 E-nr↓
卒 O↓
， O↓

↓
宋 B-nr↓
武 I-nr↓
公 E-nr↓
生 O↓
仲 B-nr↓
子 E-nr↓
， O↓
仲 B-nr↓
子 E-nr↓
生 O↓
而 O↓

古文实体识别流程

◆格式转换

- 程序所需Python包

os	re	tqdm
zhon	string	

- 依次输入上表中未安装的包按照以下格式后按回车键，待安装完成该包再继续下一个包：

```
pip install os
pip install re
pip install tqdm==4.64.0
pip install zhon==1.1.5
pip install string
```

誠 樸 篤 仁

古文实体识别流程

◆格式转换

- 打开集成开发环境（推荐使用PyCharm或Vscode等），新建一个python文件，命名为pro_ner.py输入以下内容引入相关包package

```
import os
from os import path
from os import listdir
import re
from tqdm import tqdm
from zhon import hanzi
from string import punctuation
punc = hanzi.punctuation + punctuation
```

誠 樸 篤 仁

古文实体识别流程

◆格式转换

➢Step1:将word/tag格式转换为word \t tag \n格式 （不带BIES）

注：\t表示制表符，宽度为4个空格，\n表示换行符

输入

```
惠公/nr 元妃/v 孟子/nr 。/w
```

输出

```
惠公	nr
元妃	v
孟子	nr
。	w
```

（相关代码查看GitHub库）

誠 樸 篤力 仁

古文实体识别流程

◆格式转换

➢Step2:将word \t tag格式转换为char \t tag \n （带BIES）

输入

```
惠公	nr
元妃	v
孟子	nr
。	w
```

输出

```
惠 B-nr
公 E-nr
元 O
妃 O
孟 B-nr
子 E-nr
。 O
```

（相关代码查看GitHub库）

古文实体识别流程

◆格式转换

➢Step3:调用上述两个自定义函数进行转换

```
def main（）:
    word_pos2word_seq（'data/filename.txt'）
  word_seq2char_seq（'data_seq/filename.txt'）


if __name__ == '__main__':
    main（）
```

➢Step4:运行上述代码

```
python pro_ner.py
```

誠樸勤仁

古文实体识别流程

◆训练集测试集划分

- 模型训练时需要将数据集划分为训练集和验证集，比例一般为9：1或99：1（数据较大时）。有时为了防止数据划分的偶然性，还需要进行**十折交叉验证**。
- 程序所需前置python包

os	numpy
tqdm	sklearn

（相关代码查看GitHub库）

```
pip install os
pip install numpy==1.21.5
pip install tqdm==4.64.0
pip install sklearn==1.0.2
```

古文实体识别流程

◆2 基于BERT的典籍古文实体识别

- ➢激活/创建虚拟环境

```
source activate bert或conda activate bert
```

```
conda create -r env_name python==3.7.6
```

- ➢安装所需前置python包 所需包详见requirements_env.txt

```
pip install -r requirements_env.txt
```

子文件夹，每个子文件夹内包含train.txt和test.txt.

古文实体识别流程

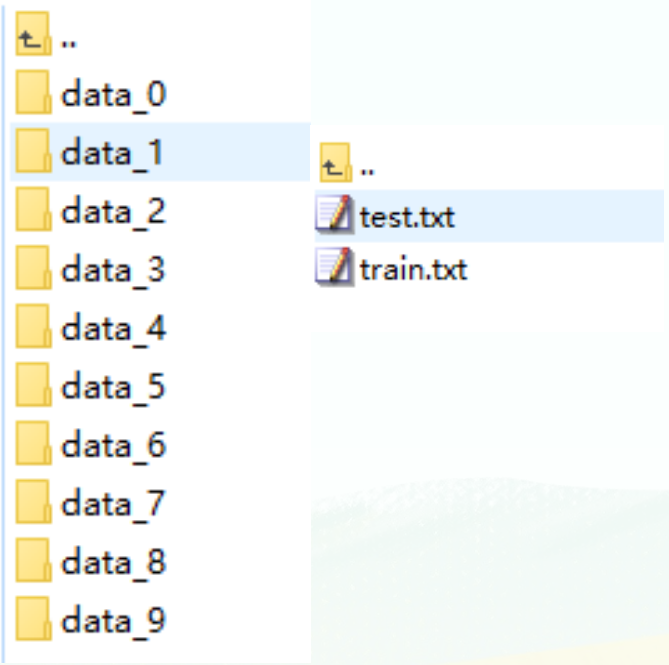
◆2 基于BERT的典籍古文实体识别

➢放入本次实验所需数据

放入位置见run.sh文件中配置：data_dir='input_folder'

其中Input_folder可以根据需要修改。

由于十折交叉验证，文件夹内包含十个data_i



古文实体识别流程

◆2 基于BERT的典籍古文实体识别

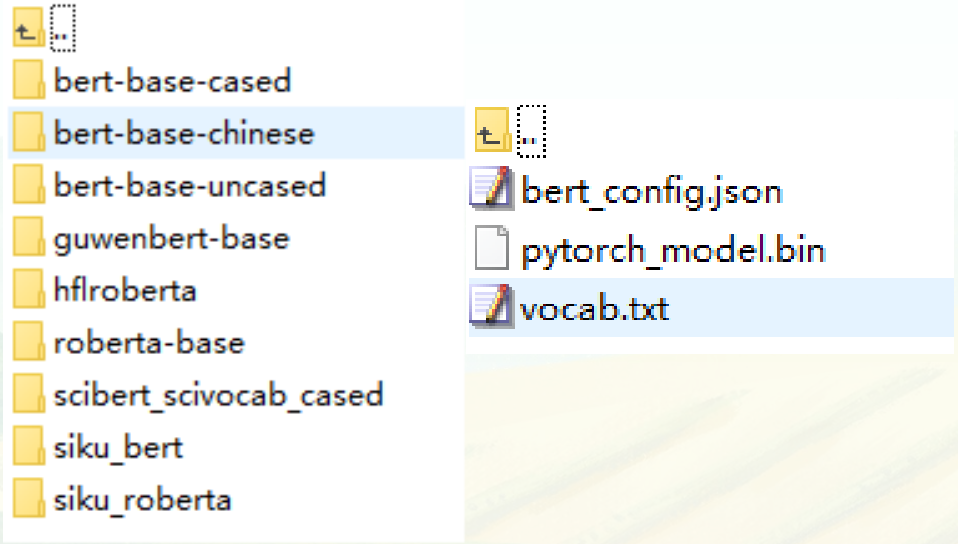
➢下载本次bert实验使用的预训练模型

放入位置见run.sh文件中配置

model_base_dir='/home/admin/pretrain_models' # 预训练模型存储路径（不含模型名）

model_dir='bert-base-chinese' # 预训练模型名

每个预训练模型需要至少如下文件：



古文实体识别流程

◆2 基于BERT的典籍古文实体识别

- ➢修改run.sh中的参数、模型路径、数据路径等

（相关代码查看GitHub库）

- ➢修改settings.py中的标签类型

```
LABELS=["X","O","B-ns","I-ns","E-ns","S-ns","B-nr","I-nr","E-nr","S-nr","B-t","I-t","E-t","S-t","[CLS]","[SEP]"]
```

注：本次识别的实体为ns、nr、t，根据实际情况修改文件

古文实体识别流程

◆2 基于BERT的典籍古文实体识别

➢运行bert模型

在终端中切换到模型所在文件夹，输入：

```
sh run.sh
```

输出文件：

Log文件，位置： log/$data_dir/$model_dir/log$i.log , 可以查看模型运行的实时输出日志。

Pid文件，位置：log/$data_dir/$model_dir /pid$i.pid ， 可以查看模型运行的线程编号。

模型结果及其他文件，位置：output/$output_dir/$model_dir/$i

古文实体识别流程

◆2 基于BERT的典籍古文实体识别

➢查看模型预测效果/结果

模型训练预测效果文件：output/$output_dir/$model_dir/$i/eval_results.txt

模型预测结果文件：output/$output_dir/$model_dir/$i/labeled_result.txt

	precision	recall	f1-score	support
nr	0.8569	0.8708	0.8638	18153
t	0.8704	0.8876	0.8789	6131
ns	0.8187	0.8179	0.8183	4816
avg / total	0.8534	0.8656	0.8594	29100

丁	B-nr	B-nr
奉	I-nr	I-nr
傳	E-nr	E-nr
丁	B-nr	B-nr
奉	E-nr	E-nr
字	O	O
承	B-nr	B-nr
淵	E-nr	E-nr
廬	B-ns	B-ns
江	E-ns	E-ns
安	B-ns	B-ns
豐	E-ns	E-ns
人	O	O
也	O	O

誠樸篤仁