**年前工作计划（1.9-2.14）**

## 1．撰写论文（1.9-2.5）

基于深度学习（应为具体的模型方法）的实体关系抽取（名字待定，英文论文）★★★

## 2.实验（1.9-2.5）

### （1）收集并分析数据（1.9-1.10） 重点★★★★★（已完成部分）

收集分析结果：

1．NYT数据集统计结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | 训练集 | 测试集 | 关系类型 | |  |
| NYT | 570089行 | 172449行 | 53种（包含None） | |  |
| 具体实体关系类型有：  NA 0  /location/neighborhood/neighborhood\_of 1  /location/fr\_region/capital 2  /location/cn\_province/capital 3  /location/in\_state/administrative\_capital 4  /base/locations/countries/states\_provinces\_within 5  /business/company/founders 6  /location/country/languages\_spoken 7  /people/person/place\_of\_birth 8  /people/deceased\_person/place\_of\_death 9  /location/it\_region/capital 10  /people/family/members 11  /location/us\_state/capital 12  /location/us\_county/county\_seat 13  /people/profession/people\_with\_this\_profession 14  /location/br\_state/capital 15  /location/in\_state/legislative\_capital 16  /sports/sports\_team/location 17  /people/person/religion 18  /location/in\_state/judicial\_capital 19  /business/company\_advisor/companies\_advised 20  /people/family/country 21  /time/event/locations 22  /business/company/place\_founded 23  /location/administrative\_division/country 24  /people/ethnicity/included\_in\_group 25 | | | | /location/mx\_state/capital 26  /location/province/capital 27  /people/person/nationality 28  /business/person/company 29  /business/shopping\_center\_owner/shopping\_centers\_owned 30  /business/company/advisors 31  /business/shopping\_center/owner 32  /people/person/ethnicity 33  /people/deceased\_person/place\_of\_burial 34  /people/ethnicity/geographic\_distribution 35  /people/person/place\_lived 36  /business/company/major\_shareholders 37  /broadcast/producer/location 38  /broadcast/content/location 39  /business/business\_location/parent\_company 40  /location/jp\_prefecture/capital 41  /film/film/featured\_film\_locations 42  /people/place\_of\_interment/interred\_here 43  /location/de\_state/capital 44  /people/person/profession 45  /business/company/locations 46  /location/country/capital 47  /location/location/contains 48  /location/country/administrative\_divisions 49  /people/person/children 50  /film/film\_location/featured\_in\_films 51  /film/film\_festival/location 52 | |
| **注：1.**train.txt及test.txt文件内容格式为：  fb\_mid\_e1, fb\_mid\_e2, e1\_name, e2\_name, relation, sentence；  2.realtion2id.txt里面存储有关系类型，共53种；  3.vec.txt存储词语向量，该向量文件从LDC Data LDC2008T19预训练得到；  4.一个实体对可能包含多个实体关系，需要特别注意 | | | | | |

2.SemEval-2010 Task8数据集统计结果：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 训练集 | 测试集 | 关系类型 |  |
| SemEval-2010 | 8000 | 2717 | 19种 |  |
| 实体关系类型（认为实体关系具有方向，故2\*9+1=19（种））：  None 0  Cause-Effect 1  Component-Whole 2  Content-Container 3  EntityDestination 4  Entity-Origin 5  Instrument-Agency 6  Member-Collection 7  Message-Topic 8  ProductProducer 9 | | | | |
| 注：1.实体关系具有方向，故认为实体类别有9\*2+1=19种  2.该数据为有监督学习数据（人工标注，质量高）  3.数据格式为：relation,e1\_start,e1\_end,e2\_start,e2\_end sentence | | | | |

3.待收集数据集。。。。。

### （2）建立一个简单的文本分类模型

文本分类模型可以基于机器学习方法，lstm，GRU，CNN等，分别在SemEval-2010 Task8数据集和NYT数据集（假设远程监督的数据基本都是对的）其中测试指标分别有：准确率P，召回率R，F1，具体结果可以填入下表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集  模型 | SemEval-2010 Task8 | | | NYT | | |
| P | R | F1 | P | R | F1 |
| SVM |  |  |  |  |  |  |
| lstm |  |  |  |  |  |  |
| gru |  |  |  |  |  |  |
| cnn |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

注：深度学习框架：pytorch,编程语言：python

机器学习方法：

以下为实验方法的具体叙述：

SVM:该方法主要是特征的选择，目前选择的大部分为词，+位置，+知网，+mention

1. 基于多分类ＳＶＭ－ＫＮＮ的实体关系抽取方法

特征：词性标注，词干提取，句法分析，谓词提取，语义角色标注

词干和词性抽取：波特词干抽取法，兰卡斯特大学提供的词性抽取算法

实体扩展特征：Cycorp提供，并采用CCG Curator进行语义角色标注

### （3）读论文，借鉴其他领域的模型，如强化学习（1.9-1.30）重点★★★

读论文结果（论述需要按照论文的格式，相关工作，言简意赅）：

### （4）收集具体的分类模型，重点记录最新论文中的创新点 重点★★★

（现有论文在前人基础上加入了什么思想，特征，效果怎样，如加入位置特征等）