“自然语言处理”实验报告

实验 2：命名实体识别

姓名：　　学号：

Email：

目录

1. 实验概述

本次实验学习使用 HMM、ME、CRF 和深度学习等不同的命名实体识别方法，并在两个不用的数据集上进行实验。实验需要用到百度 AI Studio 平台和华为云计算平台。

2. 实验目标

- 通过不同方法的结果对比，掌握不同实体识别方法的优缺点。
- 通过对不同数据集的使用，掌握命名实体识别需要的数据预处理、模型训练、模型测试和评价方法。
- 通过对百度 AI Studio 平台和华为云平台的使用，了解国产主流人工智能平台提供的学习资源和计算资源，并能使用这些平台完成特定的自然语言处理任务。
- 通过实验报告的撰写，提高分析能力、写作能力和表达能力。

3. 命名实体识别的评价

3.1 命名实体识别的评价指标

命名实体识别的评价指标有准确率(P)，召回率(R)以及 F1 值，其计算公式如下：

$$P = \frac{\text{预测结果中正确的实体数}}{\text{预测结果中所有的实体数}}$$

$$R = \frac{\text{预测结果中正确的实体数}}{\text{标准答案中所有的实体数}}$$

$$F_1 = \frac{2PR}{P+R}$$

3.2 代码实现思路

实体级命名实体识别评价过程中需要计算的中间量有三个：预测结果中的所有实体数、标准答案中的所有实体数、预测结果中正确的实体数。由于在实验过程中使用 BMES 标注法，其中预测结果中的所有实体数和标准答案中的所有实体数计算需要满足如下条件才能记为一个实体:如果实体标志为 S，则记为一个实体；如果实体以 E 开头，以 B 结尾，中间标记均为 M 且在 E~B 过程中标记中的实体类型始终不变（例如开始是 LOC，则始终是 LOC），则记为一个实体。除上述两种情况外的标记方式均不计入实体数目。

在计算正确预测的实体数时，只有满足如下条件记为正确：当预测结果和标准答案都出现相同实体对应的 B 标记（如 B-LOC）时表示开始，而中间的 M 标记应该保持实体类型不变，到最终的 E 标记为止，这过程中的每一个值都应该是属于同一种实体的标记，且这一过程中的每一个值在预测结果和标准答案结果中都应该相等。计算正确预测数流程图如下：

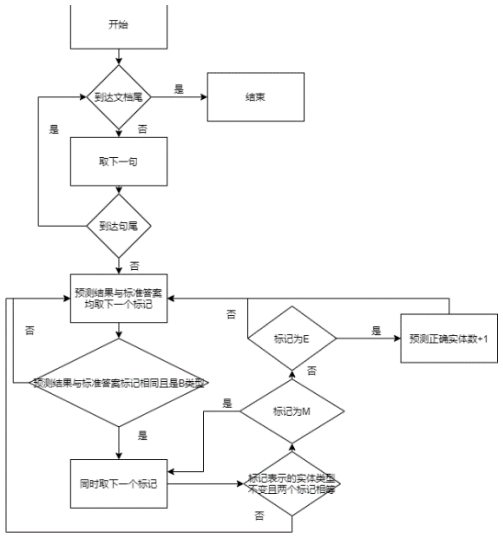

图 1 正确预测实体数计算流程图

3.3 评价程序调用方法

实验过程中将此程序封装为函数，函数名为 score，函数形参为 pred_lists，standard_lists 与 tag_id，其中 pred_lists 表示使用模型预测出的 tag 序列，其格式为二维输出，其中每一行表示对每一个句子的预测，而每一行中的每一列表示对句子中的每一个字的预测结果，分别是 BMES 中之一与对应的实体类型或是 O 表示不是实体的一部分；standard_lists 表示标准结果中的 tag 序列，其格式与 pred_lists 相同，每一位的含义与格式也相同；tag_id 表示训练数据集中出现过的每一种实体，例如 LOC，由于在使用模型预测时可能的预测结果中的实体一定是在训练数据集中出现过的，因此所有的不包含在训练数据集中出现的实体中的实体均视为 O。函数的返回值有两部分，第一部分是所有类型的实体对应的评价结果，以字典形式存储，其中每一个 key 对应一种实体，每一个 key 都是一个 dic，其中存储评价需要的准确率，召回率，F1；第二部分是整体的准确率，召回率和 F1，也是以字典形式存储。

在调用过程中只需要运行 notebook 中对应的 cell 即可。

3.4 HMM 评价方式结果差别分析

对 HMM 模型在 ner_char_data 目录下的 test.txt 文件上的识别结果分别按词级别和实体级别进行评价时评价结果的分部分和整体 P、R、F1 值分别如下：

	precision	recall	f1-score
B-NAME	0.9800	0.8750	0.9245
M-NAME	0.9459	0.8537	0.8974
E-NAME	0.9000	0.8036	0.8491
O	0.9568	0.9177	0.9369
B-PRO	0.5581	0.7273	0.6316
E-PRO	0.6512	0.8485	0.7368
B-EDU	0.9000	0.9643	0.9310
E-EDU	0.9167	0.9821	0.9483
B-TITLE	0.8811	0.8925	0.8867
M-TITLE	0.9038	0.8751	0.8892
E-TITLE	0.9514	0.9637	0.9575
B-ORG	0.8422	0.8879	0.8644
M-ORG	0.9002	0.9327	0.9162
E-ORG	0.8262	0.8680	0.8466
B-CONT	0.9655	1.0000	0.9825
M-CONT	0.9815	1.0000	0.9907
E-CONT	0.9655	1.0000	0.9825
M-EDU	0.9348	0.9609	0.9477
B-RACE	1.0000	0.9286	0.9630
E-RACE	1.0000	0.9286	0.9630
B-LOC	0.3333	0.3333	0.3333
M-LOC	0.5833	0.3333	0.4242
E-LOC	0.5000	0.5000	0.5000
M-PRO	0.4490	0.6471	0.5301
avg/total	0.9149	0.9122	0.9130

图 2 词级别评价

	precision	recall	f1-score
NAME	0.8491	0.8036	0.8257
CONT	0.9655	1.0000	0.9825
RACE	0.8667	0.9286	0.8966
TITLE	0.8747	0.8860	0.8803
EDU	0.8917	0.9554	0.9224
ORG	0.7479	0.7884	0.7676
PRO	0.5581	0.7273	0.6316
LOC	0.3333	0.3333	0.3333
total	0.8219	0.8491	0.8352

图 3 实体级别评价

通过分析命名实体识别的词级别和实体级别评价结果可以发现，在各个类别的命名实体识别中以及整体识别指标上实体级别的评价都略低于词级别的评价，主要原因是进行词级别评价时只需要一个词的标记正确则正确标记数加一，而实体级别评价时存在一个实体中部分词标记正确，部分词标记错误的情况，在实体级别的评价中这并不算正确的数量。例如一个实体的标准划分标签为[B-NAME,M-NAME,E-NAME]，而预测的划分标签为[B-NAME,E-NAME,E-NAME]，则对于词级别的评价来说，由于有一个 B-NAME 预测正确，因此 B-NAME 的正确预测数量加一；而对于实体级别的评价来说，由于这个实体预测错误，因此对于正确预测数量不造成影响。从这个例子可以看出实体级别的评价比词级别的评价 P、R、F1 值都略低的原因。

4.基于最大熵模型的实体识别

4.1 最大熵模型原理

最大熵模型属于对数线性分类模型，在损失函数优化过程中使用了与支持向量机类似的凸优化技术，其主要原理如下：

首先引入熵的概念，在信息论中，事件的不确定性越大，熵就越大，具体来说，熵的定义如下：$H(X)-\sum_{i=1}^{n}p_i\,logp_i$，其中 n 代表 X 的 n 种不同的离散取值，而p_i代表了 X 取值为 i 的概率。容易推广到多个变量的联合熵，这里给出两个变量的联合熵表达式：

$$H(X,Y)=-\sum_{x_i\in X}\sum_{y_i\in Y}p\,(x_i,y_i)logp(x_i,y_i)$$

定义了联合熵之后，可以得到条件熵表达式，条件熵类似于条件改了，度量了在知道 X 之后 Y 的不确定性，条件熵表达式如下：

$$H(X|Y)=-\sum_{x_i\in X}\sum_{y_i\in Y}p\,(x_i,y_i)logp(y_i|x_i)=\sum_{j=1}^{n}p\,(x_j)H(Y|x_j)$$

最大熵模型假设分类模型是一个条件分布概率$p(Y|X)$，其中 X 为特征，Y 为输出。

给定训练集$(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)},\ldots,(x^{(m)},y^{(m)}))$，其中 x 是 n 维特征向量，y 为类别输出，目标是使用最大熵模型获得一个最好的分类模型。在给定训练集的情况下，我们可以得到总体联合分布$P(X,Y)$的经验分布$\widehat{P}\,(X,Y)$和边缘分布$P(X)$的经验分布$\widehat{P}\,(X)$。

使用特征函数$f(x,y)$描述输入 x 与输出 y 之间的关系的时候定义如下：

$$f(x,y)=\begin{cases}1 & \text{x 与 y 满足某个关系}\\ 0 & \text{否则}\end{cases}$$

特征函数$f(x,y)$关于经验分布$\widehat{P}\,(X,Y)$的期望值表示如下：

$$E_{\widehat{P}}(f)=\sum_{x,y}\widehat{P}\,(x,y)f(x,y)$$

特征函数关于条件分布$P(Y|X)$和经验分布$\widehat{P}\,(X)$的期望值表示如下：

$$E_p(f)=\sum_{x,y}\widehat{P}\,(x)P(y|x)f(x,y)$$

如果模型可以从训练集中学习，假设这两个期望相等，这就是最大熵模型学习的约束，假设有 M 个特征函数，则有 M 个约束条件。

这样我们就得到了最大熵模型的定义如下：

假设满足所有约束条件的模型集合为：$E_{\hat{P}}(f_i) = E_p(f_i), i = 1,2,\dots M$

定义在条件概率分布上的条件熵为$H(P) = -\sum_{x,y} P\,(y|x) log P(y|x)$。我们的目标就是得到 H(P)最大时对应的条件概率。

通过将凸优化问题转化为无约束优化问题并进行求最值操作之后最终可以得到我们要求的结果表达式如下：

$$P_w(y|x) = \frac{1}{Z_w(x)} exp(\sum_{i=1}^{M} w_i\, f_i(x,y))$$

其中$Z_w(x)$为规范化因子，定义为：$Z_w(x) = \sum_y e\,xp(\sum_{i=1}^{M} w_i\, f_i(x,y))$

得到求解结果的形式之后可以采用改进的迭代尺度法（IIS）求解。

4.2 基于最大熵模型的实体识别系统实现思路

最大熵模型中最重要的就是定义需要使用的特征，在本实验中使用的特征是需要预测的词的前后各两个词组成的 5 个字大小的窗口作为的特征，分别使用 5 个单字、双字以及三字特征作为计算的特征。由于需要开一个 5 个字大小的窗口，因此在句首和句尾分别定义了两个句首和两个句尾标记。经过测试，在本实验中使用更大的窗口并不能带来性能的提升。

在实现过程中主要基于最大熵的表达式，对于每一个字都抽取其前后文字组成的窗口特征，并使用 DictVectorizer()将文字型特征转换为数字型离散特征便于后续训练。

在训练过程中使用 sklearn 中的 LogisticRegression 函数进行训练，使用的训练方式为“saga”，这一训练方式对于大型训练集训练速度相较于其他几种训练方式较有优势；在测试过程中使用的迭代轮数有 50,100,200，通过函数的训练情况可以发现三者都没有收敛，但是相对来说 50 代的时候训练结果和后两者相比并没有太大的差别，因此最终决定使用的迭代轮数为 50 代。

4.3 最大熵模型实现效果

最大熵模型在 ner_char_data 目录下的 train.txt 文件训练模型，在 test.txt 文件上进行测试后测试结果如下：

	precision	recall	f1-score
B-NAME	0.8140	0.9375	0.8714
M-NAME	0.9878	0.9878	0.9878
E-NAME	1.0000	0.9911	0.9955
O	0.9666	0.9482	0.9573
B-PRO	0.7600	0.5758	0.6552
E-PRO	0.8611	0.9394	0.8986
B-EDU	0.9646	0.9732	0.9689
E-EDU	0.9909	0.9732	0.9820
B-TITLE	0.9316	0.9171	0.9243
M-TITLE	0.9371	0.8829	0.9092
E-TITLE	0.9909	0.9870	0.9890
B-ORG	0.9359	0.9241	0.9299
M-ORG	0.9127	0.9646	0.9379
E-ORG	0.9216	0.8933	0.9073
B-CONT	0.9655	1.0000	0.9825
M-CONT	1.0000	1.0000	1.0000
E-CONT	1.0000	1.0000	1.0000
M-EDU	0.9714	0.9497	0.9605
B-RACE	1.0000	1.0000	1.0000
E-RACE	1.0000	1.0000	1.0000
B-LOC	1.0000	0.3333	0.5000
M-LOC	0.8182	0.4286	0.5625
E-LOC	1.0000	0.8333	0.9091
M-PRO	0.6329	0.7353	0.6803
avg/total	0.9414	0.9406	0.9404

图 4 词级别评价

	precision	recall	f1-score
NAME	0.8062	0.9286	0.8631
CONT	0.9655	1.0000	0.9825
RACE	1.0000	1.0000	1.0000
TITLE	0.9026	0.8886	0.8956
EDU	0.9469	0.9554	0.9511
ORG	0.7949	0.7848	0.7898
PRO	0.8636	0.5758	0.6909
LOC	0.0000	0.0000	0.0000
total	0.8619	0.8540	0.8579

图 5 实体级别评价

5. HMM、ME 与 CRF 的效果对比

5.1 使用新训练集训练方式

在使用新数据集训练的时候大体的思路和使用 char 数据集类似，都是使用一个训练集训练之后使用测试集测试性能。在使用 ner_clue_data 目录中的文件进行训练的过程中首先读取文件，读取文件的时候处理格式参考 ner_char_data 中的处理方式，对每个句子中的每个字都给一个标记，不是实体一部分的给 O 标记，其余按照在实体中的位置给出标记。

5.2 模型结果差异分析

使用三种模型的测试结果如下表所示：

	HMM			ME			CRF		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
company	0.5049	0.5476	0.5254	0.6052	0.4947	0.5444	0.7955	0.7407	0.7671
name	0.5573	0.6065	0.5808	0.6388	0.5097	0.567	0.8082	0.7247	0.7642
game	0.6291	0.7186	0.6709	0.7021	0.6949	0.6985	0.8161	0.8271	0.8215
Organization	0.5182	0.5422	0.53	0.6916	0.6049	0.6453	0.8062	0.7138	0.7572
movie	0.533	0.6424	0.5826	0.4758	0.3907	0.4291	0.6818	0.6954	0.6885
position	0.6398	0.6605	0.65	0.8182	0.6443	0.7209	0.8191	0.7113	0.7614
address	0.3582	0.3727	0.3653	0.342	0.2466	0.2866	0.5932	0.4692	0.524
government	0.5052	0.5951	0.5465	0.5404	0.4332	0.4809	0.7846	0.7814	0.783
scene	0.4417	0.4354	0.4386	0.4731	0.2105	0.2914	0.6433	0.4833	0.5627
book	0.6267	0.4026	0.4901	0.48	0.3896	0.4301	0.7869	0.6234	0.6957
total	0.5278	0.5605	0.5448	0.6107	0.4857	0.5411	0.7679	0.6839	0.7235

分析测试结果可以发现，ME 和 HMM 的性能指标相近，而 CRF 相对来说比前两种算法略好一些。分析原因如下：显然最大熵模型还有很大的性能提升空间，但是本实验中受限于使用的特征比较简单也比较单一，因此性能指标相对来说并不是特别突出。接下来分析三种模型各自的优劣：

HMM 模型将标注看作马尔可夫链，一阶马尔可夫链式针对相邻标注的关系进行建模，其中每个标记对应一个概率函数。HMM 模型的这个假设前提在比较小的数据集上是合适的，但实际上在大量真实语料中观察序列更多的是以一种多重的交互特征形式表现，观察元素之间广泛存在长程相关性。在命名实体识别的任务中，由于实体本身结构所具有的复杂性，利用简单的特征函数往往无法涵盖所有的特性， HMM 的假设前提使得它无法使用复杂特征(它无法使用多于一个标记的特征)。

最大熵模型理论上可以使用任意复杂的相关特性，但是每个词都是单独分类的，没有使用两个词标记之间的关系，这与 HMM 使用了前文信息是不一样的。最大熵的优点在于其可以根据任务的要求和性能的要求灵活设置约束条件；同时可以自然解决统计模型中的参数平滑问题。而其缺点在于计算开销比较大，时空开销也较大。同时数据稀疏问题比较严重。

CRF 模型在给定给定了观察序列的情况下，对整个的序列的联合概率有一个统一的指数模型。相比于 HMM 来说能够更好地使用上下文信息。CRF 模型的优点：首先，CRF 模型由于其自身在结合多种特征方面的优势和避免了标记偏置问题。其次，

CRF 对特征的融合能力比较强，对于实例较小的 ME 来说，CRF 的识别效果明显高于 ME 的识别结果。CRF 模型的不足：首先，通过对基于 CRF 的结合多种特征的方法识别英语命名实体的分析，发现在使用 CRF 方法的过程中，特征的选择和优化是影响结果的关键因素，特征选择问题的好与坏，直接决定了系统性能的高低。

6. 在华为云上的计算资源进行命名实体识别

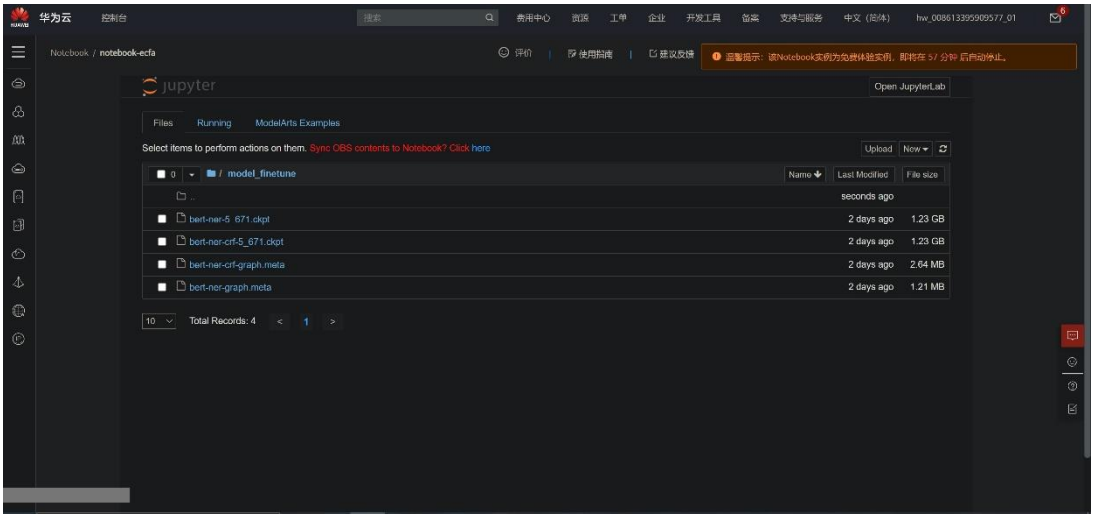

图 6 所有训练结果

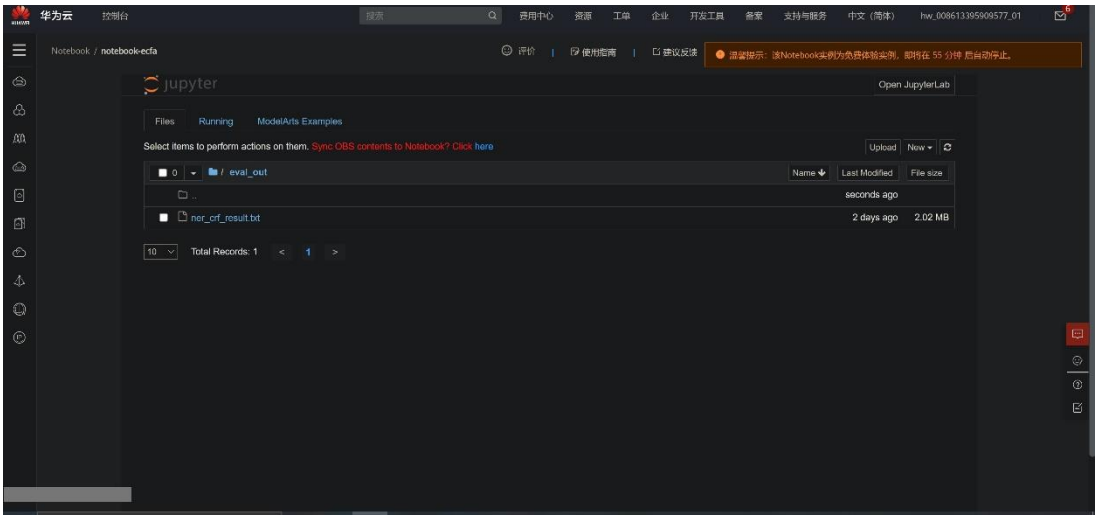

图 7 最终测试结果文件

图 8 最终测试文件部分内容

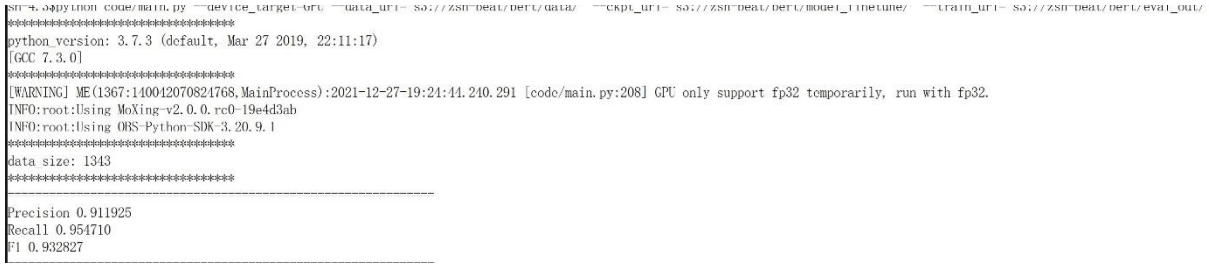

图 9 微调过程测试结果部分

7. 实验的收获和体会

经过本次实验对于 HMM、ME、CRF 用于命名实体识别的方式和实现有了更加深入的了解，对于命名实体识别的原理也有了一定的体会。同时通过使用华为云资源训练基于 bert 和 CRF 的命名实体识别方式，通过体验当前精度最高的命名实体方式，加深了对相关理论的理解。

参考文献

[1] https://blog.csdn.net/sddxyf6/article/details/10174803

[2] https://www.cnblogs.com/pinard/p/6093948.html

[3] 宗成庆.统计自然语言处理（第二版）[M].北京：清华大学出版社，2008.05