汉语词义消岐实验报告

霍华荣 1301210680、胡腾 1301210664

2014年4月2日

1. 实验目的

1.1 问题描述

通过试验不同的机器学习算法，根据训练集的标注语料，完成汉语词义消岐（WSD）的任务，并比较不同机器学习算法在此问题中的效果。

1.2 相关语料

训练语料共2686个句子，包含标注词语Type 40个；测试语料共935个句子，共含待消岐词语Type 40个。

1.3 评测方法

微平均：

$$MicroAve = \sum_{i=1}^{N} m_i \Big/ \sum_{i=1}^{N} n_i$$

宏平均：

$$MacroAve = \sum_{i=1}^{N} p_i \Big/ N \,,\ p_i = m_i / n_i$$

2. 特征提取

2.1 标记特征：

- 目标词前后指定窗口内的单词、嵌套短语、对应词性; 特征带有位置和类别标签。

2.2 词袋特征：

- 目标词前后指定窗口内的实词（本文主要考虑了名词、动词两种）,不带位置标记。

2.1 标记特征

- W-i：左边第i个单词（W指单词），
- Wi: 右边第i 个单词
- T-j：左边第j个单词POS（T指词性标签）
- Tj: 右边第j个单词POS
- NULL_HEAD: 左边指定位置属性不存在的空缺值
- NULL_TAIL：右边指定位置属性不存在的空缺值
- 最后一个为所属的意思，集训练结果
- PW：被嵌入的短语
- PT：短语的POS

- 其中 Wnd_l <= i <= Wnd_r, i != 0；Wnd_l <= j <= Wnd_r, j!= 0

2.2 词袋特征

- 参数CWnd：
- 表示词袋窗口大小（目标词左边、右边距离在CWnd以内的实词）。

特征提取参数训练

- 特征提取中共有三个参数：
  - CWnd
  - Wnd_l
  - Wnd_r

- 应用朴素贝叶斯模型，根据最大似然估计的方法训练

3. 实验方法

- 朴素贝叶斯
- 神经网络
- 最大熵模型
- 支持向量机
- 决策树

3.1 朴素贝叶斯

- 验证方法：

- 采用 4-fold 随机交叉验证，从训练集每个单词每个词义的所有样本中随机抽取1/4的样本组成调试集，剩余3/4样本为训练集，进行训练并测试;
- 重复抽样、训练和测试20次，每个词得到一样平均正确率，所有词再得到一个正确率的宏平均值Macro AVG，以该值为标准进行模型优化。

- Add λ 平滑值粗调：
- 根据一组简单的特征（CWnd = 0, Wnd_l = 2, Wnd_r = 2）初步确定平滑值λ。

λ Gross Tuning	
λ	Macro AVG
0.5	0.564900153
0.1	0.686026405
0.01	0.714557336
0.001	**0.730018373**
0.0001	0.730041184

- 优化特征筛选参数:

CWnd	Macro AVG
1	0.509908292
2	0.55382535
3	0.557113308
4	0.565982526
5	0.574454965
6	0.575142773
7	**0.580318117**
8	0.578937699
9	0.574370971
10	0.578529258

CWnd-Wnd_l- Wnd_r	Macro AVG
7-1-1	0.717748511
7-1-2	0.723734342
7-1-3	0.719334134
7-2-1	0.730982496
7-2-2	0.737257622
7-2-3	0.737628263
7-3-1	0.733481717
7-3-2	**0.745448969**
7-3-3	0.742495853
7-4-1	0.718212898

- Add λ 平滑值微调
  - CWnd-Wnd_l- Wnd_r = 7-3-2

λ	Macro AVG
0.001	0.745448969
0.0005	0.744985585
0.0001	**0.755031757**
0.00005	0.752713891
0.00001	0.752613186

- Test 集运行结果
  - CWnd-Wnd_l- Wnd_r = 7-3-2， λ =0.0001

- Micro AVG：0.728342
- Macro AVG：0.767368

3.2 神经网络

- 参数设置：
  - 输入结点数：自适应调整为特征数
  - 输出结点数：自适应调整为类的个数
  - 隐藏层数：1
  - 隐藏层结点数：15
  - 迭代次数： 40
  - 学习率： 0.5
- Test 集运行结果：
  - Micro AVG：0.712299
  - Macro AVG：0.753614

3.3 最大熵模型

- 参数：
  - 迭代算法：iis
  - 迭代次数：80
- 测试结果：
  - MicroAVG: 0.699465
  - MacroAVG: 0.728432

3.4 支持向量机

- 参数
  - 核函数：rbf
  - C: 1000.0
  - gamma: 0.0001
- 参数C, gamma选择算法：
  1. 固定C，以10的倍数调整gamma，选择预测评价最高的gamma值；
  2. 固定gamma，以10的倍数调整C，选择预测评价最高的C；
  3. 重复1，2，直到 C和gamma值收敛。
- 测试结果
  - MicroAVG: 0.727273
  - MacroAVG: 0.762196

两种特征值数值化处理方法

1. 固定特征名的个数，将特征值映射到唯一的标号。
   - 旧特征：{W-2: "而", T-2: "c", W-1: "钻研", T-1: "v", W1: "理论", T1: "n", W2: "，", T2: "w"}
   - 新特征： {W-2: 0, T-2: 1, W-1: 3, T-1: 4, W1: 5, T1: 6, W2: 7, T2: 8}
2. 将特征名和特征值一起作为新的特征名，新的特征值均为1。
   - 旧特征：{W-2: "而", T-2: "c", W-1: "钻研", T-1: "v", W1: "理论", T1: "n", W2: "，", T2: "w"}
   - 新特征：{W-2=而: 1, T-2=c: 1, W-1=钻研: 1, T-1=v: 1, W1=理论: 1, T1=n: 1, W2=，: 1, T2=w: 1}

- 对于以上两种数值化处理方法，均应用以下8个特征：W-2, T-2, W-1, T-1, W1, T1, W2, T2, NamedEntity, NamedEntityType。
  1. MicroAVG: 0.485561　MacroAVG: 0.552010
  2. MicroAVG: 0.713369　MacroAVG: 0.739836

3.5 决策树

- 参数
  - 熵阀值：0.05
  - 深度阀值：100
  - 节点元素个数阀值：10
- 损失函数
  - 损失函数值 = 按照此特征分类错误实例个数 / 本次待分类实例总数
- 各层特征选择
  - 选择损失函数值最小的的特征。
- 结果
  - MicroAVG: 0.658824
  - MacroAVG: 0.700084

4. 分析

- 不同特征、模型对比

特征模式	7-3-2		7-1-1	
正确率	微平均	宏平均	微平均	宏平均
朴素贝叶斯	0.728342	0.767368	0.732620	0.767677
神经网络	0.712299	0.753614	0.735829	0.781223
最大熵模型	0.699465	0.728432	0.727273	0.757359
支持向量机	0.727273	0.762196	0.734759	0.772213
决策树	0.658824	0.700084	0.682353	0.719701

与优秀实现比较

- 何径舟等研究了基于特征选择和最大熵模型的汉语词义消岐[1]，应用特征模板和特征自动选择机制，基于最大熵模型，测出最高正确率为微平均0.7476, 宏平均0.7788。
- 本实验的最大熵模型正确率大概小0.02，说明本实验，尤其是特征选择部分，还有一定的改进空间。

- [1] 何径舟,王厚峰. 基于特征选择和最大熵模型的汉语词义消歧. 软件学报, 2010, 21(6): 1287-1295.

多分类器集成

- 吴云芳等研究了多分类器集成的汉语词义消岐研究[2]，认为使用乘法、均值、最大值的集成方法均表现出良好的分类性能，3种方法的消岐准确率均高于单一分类器。
- 因此，本实验后续还可以应用不同集成方法将上述5中消岐模型进行整合，构建更优集成模型。

- [2] 吴云芳,王淼,金澎,俞士汶. 多分类器集成的汉语词义消歧研究. 计算机研究与发展, 2008,45(8): 1354-1361.

5. 结论

- 汉语词义消岐是一个分类任务。基本上机器学习中的分类算法都能够应用到此问题中，并且都能够得到不错的效果。如果只看中正确率，那么神经网络模型具有最高宏平均；如果兼顾效率和正确率，那么支持向量机模型无疑更加实用。
- 无论应用什么模型，最后预测的准确率均极大的与选择的特征相关。因此应该将更多的时间用在优化特征选择上。
- 75%左右的词义消岐正确率并不高，离实际应用的要求还有较大的距离，更优秀的算法和算法组合有待提出。