词义消岐(WSD)实验报告

霍华荣 1301210680、胡腾 1301210664

2014年4月2日

# 实验目的

### 1.1 问题描述

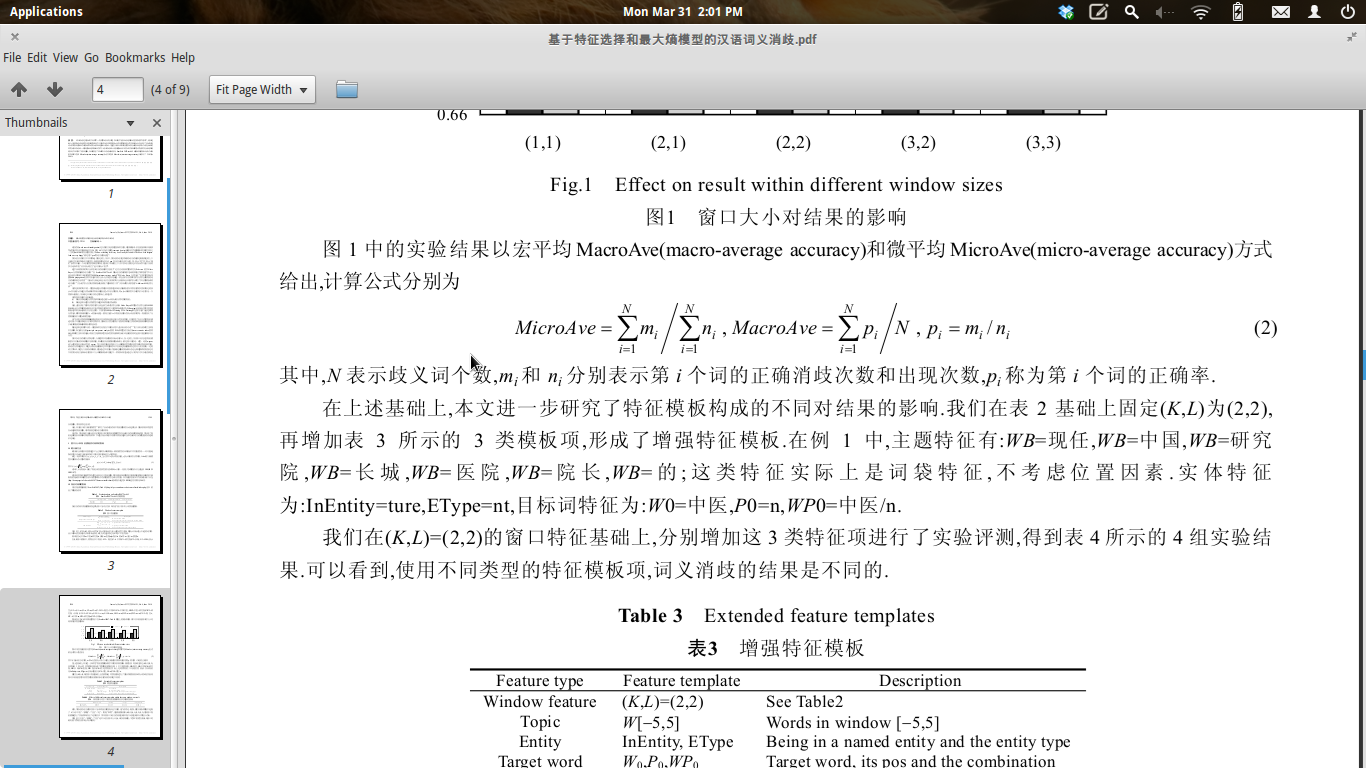
通过试验不同的机器学习算法，根据训练集的标注语料，完成中文词义消岐（WSD）的任务，并比较不同机器学习算法在此问题中的效果。

### 1.2 相关语料

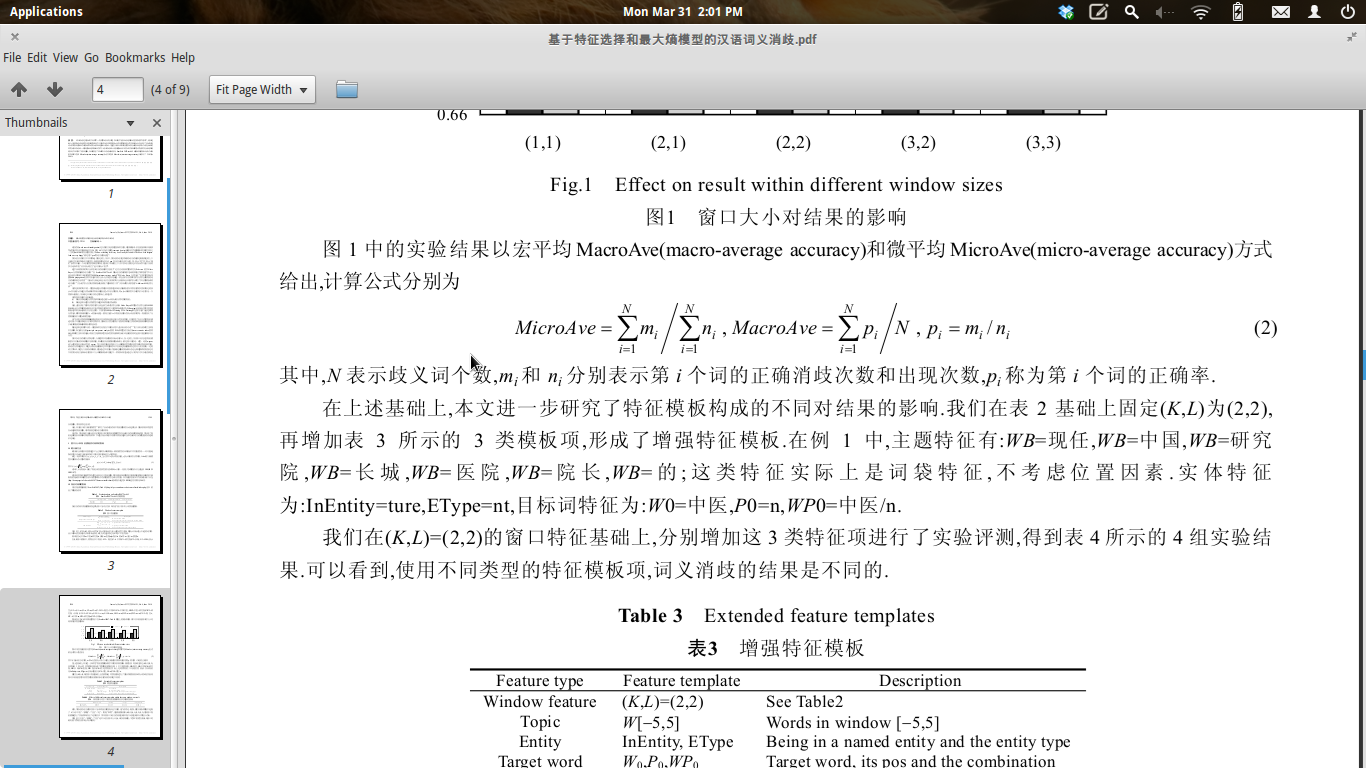
训练语料共2686个句子，包含标注词语Type 40个；测试语料共935个句子，共含待消岐词语Type 40个。

### 1.3 评测方法

###### 微平均：



###### 宏平均：



# 2. 特征提取

# 3. 实验方法

### 3.1 神经网络

### 3.2 朴素贝叶斯

### 3.3 最大熵模型

###### 参数

迭代算法：iis

迭代次数：80

###### 测试结果

Total: 935 Finished: 935 Correct: 680 MicroAVG: 0.727273 MacroAVG: 0.757359

### 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。

###### 特征值数值化处理

从语料中提取的特征的值都是以词语，也就是字符串的形式。但是对于支持向量机模型而言，特征的值必须为数值型。因此需要对特征的值进行数值化处理。

共有两种的数值化处理方法：

1. 固定特征名的个数，将特征值映射到唯一的标号。

例如：

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}

1. 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。

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。得出的测试结果为：

处理方法1： Total: 935 Finished: 935 Correct: 454 MicroAVG: 0.485561 MacroAVG: 0.552010

处理方法2：Total: 935 Finished: 935 Correct: 667 MicroAVG: 0.713369 MacroAVG: 0.739836

从上述结果看出，处理方法2效果明显更好。依次后续SVM算法均以此方法进行特征数值化处理。

###### 测试结果

Total: 935 Finished: 935 Correct: 687 MicroAVG: 0.734759 MacroAVG: 0.772213

### 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

Total: 935 Finished: 935 Correct: 638 MicroAVG: 0.682353 MacroAVG: 0.719701

# 4. 分析

# 5. 结论

# 6. 试验环境

操作系统: Elementary OS Luna 64 bits (基于 Ubuntu 12.04)

编程语言: Python 2.7.3

工具模块: NLTK, Scikit-learn

CPU: Intel(R) Core(TM)2 Duo 2.20GHz

内存: 2G