基于深度学习的微博情感分析

梁军 zhengdaxg@163.com
指导老师：柴玉梅 昝红英
郑州大学自然语言处理实验室
http://nlp.zzu.edu.cn
2014年10月18日

报告提纲

1. 背景介绍
2. 微博情感分析模型
   - 2.1 词语的表示
   - 2.2 递归自编码(Recursive AutoEncoder)
   - 2.3 基于RAE的情感极性转移模型
   - 2.4 情感分析模型算法
3. 实验结果分析
4. 总结

报告提纲

1. 背景介绍
2. 微博情感分析模型
   - 2.1 词语的表示
   - 2.2 递归自编码(Recursive AutoEncoder)
   - 2.3 基于RAE的情感极性转移模型
   - 2.4 情感分析模型算法
3. 实验结果分析
4. 总结

1. 背景介绍

情感分析

文本情感分析(sentiment analysis)，又称为意见挖掘，是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。其中，主观情感可以是他们的判断或者评价，他们的情绪状态，或者有意传递的情感信息。因此，情感分析的一个主要任务就是情感倾向性的判断。

情感倾向性判断方法 🙂 or 🙁 ?

1. 基于情感词典的方法
2. 转化为分类问题使用机器学习方法

报告提纲

1. 背景介绍
2. 微博情感分析模型
   - 2.1 词语的表示
   - 2.2 递归自编码(Recursive AutoEncoder)
   - 2.3 基于RAE的情感极性转移模型
   - 2.4 情感分析模型算法
3. 实验结果分析
4. 总结

2. 微博情感分析模型

本文提出一个基于RAE的情感极性转移模型，该模型首先将文本数据转化为低维实数向量表示，建立表示文本特征的矩阵，然后将其作为基于RAE的情感极性转移模型的输入，最后使用LBFGS算法迭代生成模型。

报告提纲

1. 背景介绍
2. 微博情感分析模型
   - 2.1 词语的表示
   - 2.2 递归自编码(Recursive AutoEncoder)
   - 2.3 基于RAE的情感极性转移模型
   - 2.4 情感分析模型算法
3. 实验结果分析
4. 总结

2.1 词语的表示(1/3)

- 传统的词语表示方法
  - 向量空间表示方法：将单词作为原子符号，其中词语通常被表示成由一串0、1代码组成的向量：[0 0 0 0 0 0 0 1 0 0 0]
  - 缺点：

    维度高：20K（语音）、50K（PTB）、13M（Google 1T）

    无法体现相关性：酒店[0 0 0 1 0 0 0 0 0] & 宾馆[0 0 0 0 0 1 0 0 0]=0

2.1 词语的表示(2/3)

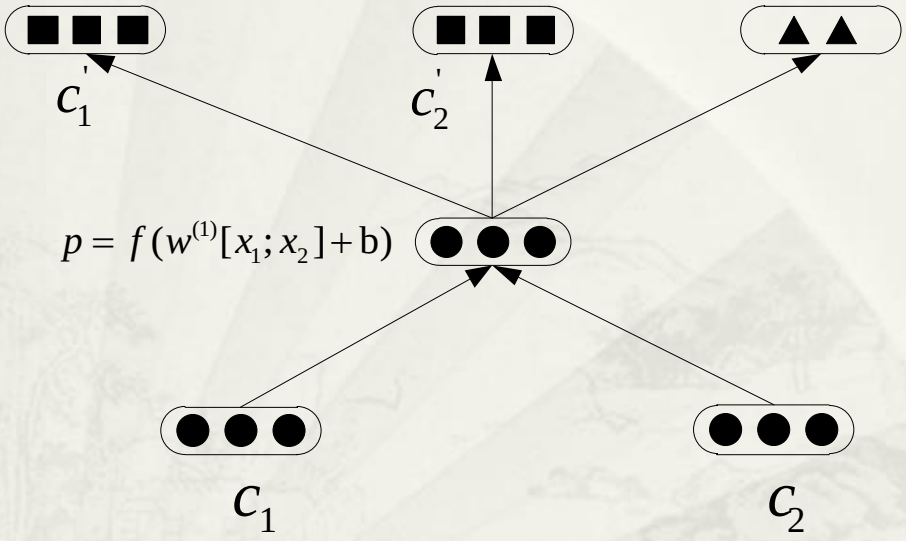
- 词向量表示方法
  - 在聚类模型中提出使用分布式方法表示词语（distribution representation）
    - Latent Semantic Analysis (LSA/LSI), Random projections
    - Latent Dirichlet Analysis(LDA), HMM clustering
  - Neural word embeddings as a distributed representation
    - Combine vector space semantics with the prediction of probabilistic models (Bengio et al. 2003, Collobert & Weston 2008, Turian et al. 2010)
    - 表示方法：酒店[0.357 0.159 -0.526 0.610]

2.1 词语的表示(3/3)

- 自编码（AutoEncoder）

$$s(p;\theta)=\mu(w^{(3)}p+b^{(3)})$$

例如“深度”的向量表示为[0.3 0.1 0.6]^T，“学习”的向量表示为[0.2 0.5 0.7]^T，那么该如何表示“深度 学习”呢？假定“深度 学习”是“深度”、“学习”的父节点p，“深度”是第一个子节点c_1，“学习”是第二个子节点c_2，那么p可由函数f从c_1、c_2映射得到：



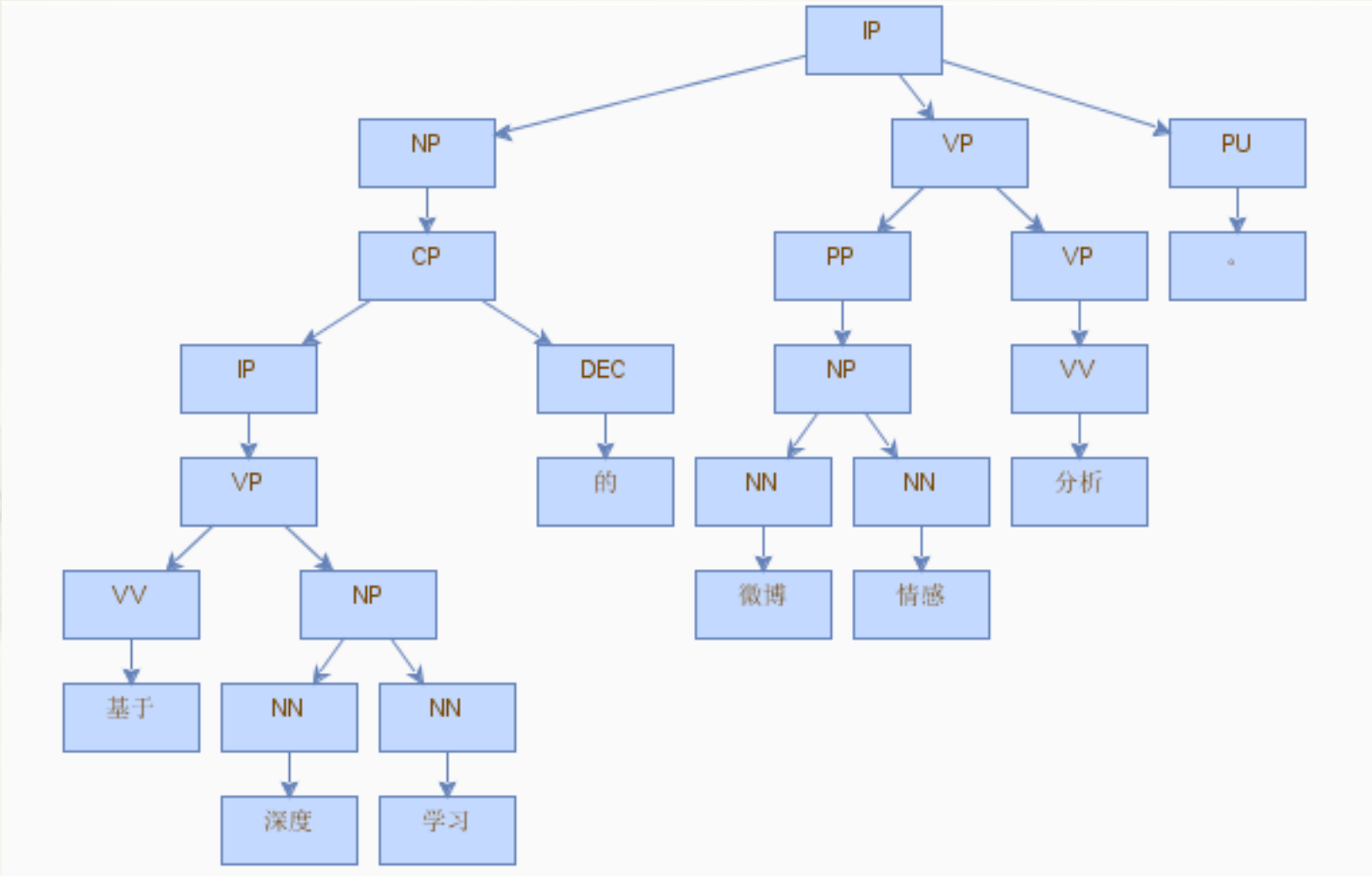
$$p=f(w^{(1)}[c_1;c_2]+b^{(1)})$$

$$E_{rec}=\|c_1-c_1'\|^2+\|c_2-c_2'\|^2$$

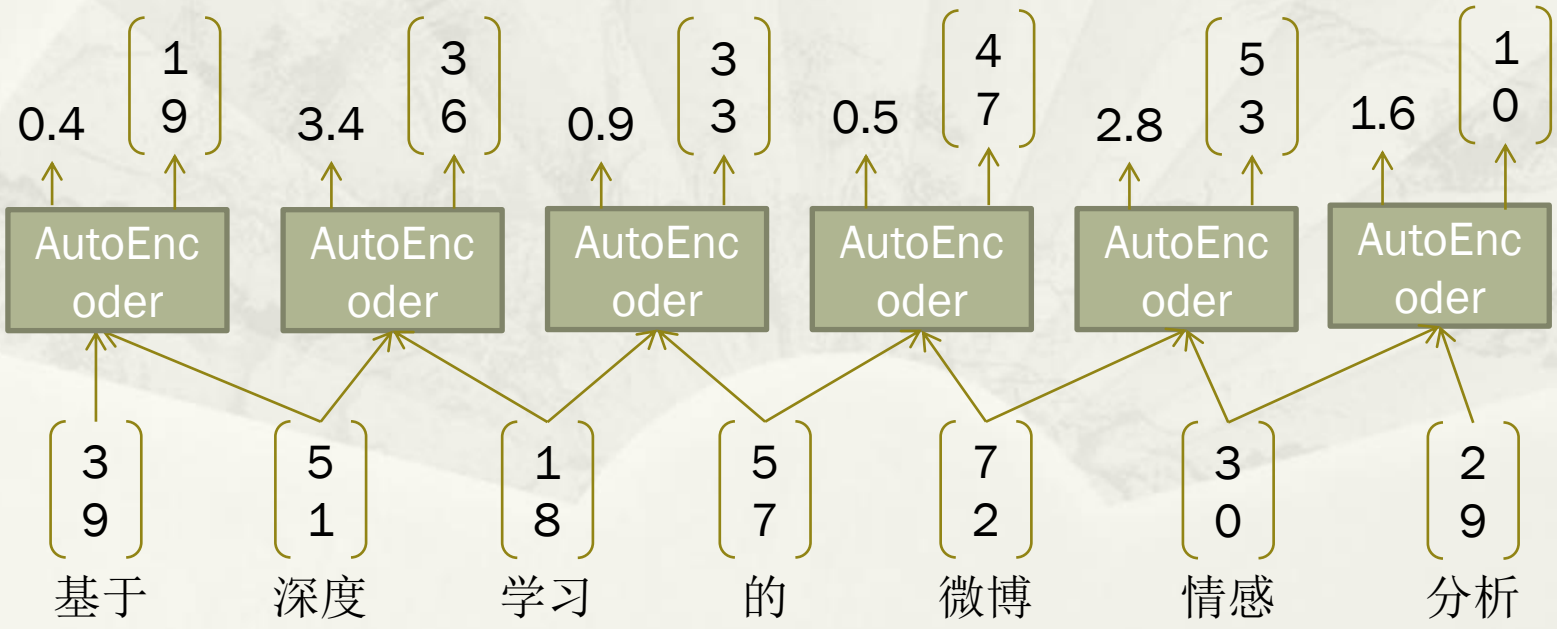
报告提纲

1. 背景介绍
2. 微博情感分析模型
   - 2.1 词语的表示
   - 2.2 递归自编码(Recursive AutoEncoder)
   - 2.3 基于RAE的情感极性转移模型
   - 2.4 情感分析模型算法
3. 实验结果分析
4. 总结

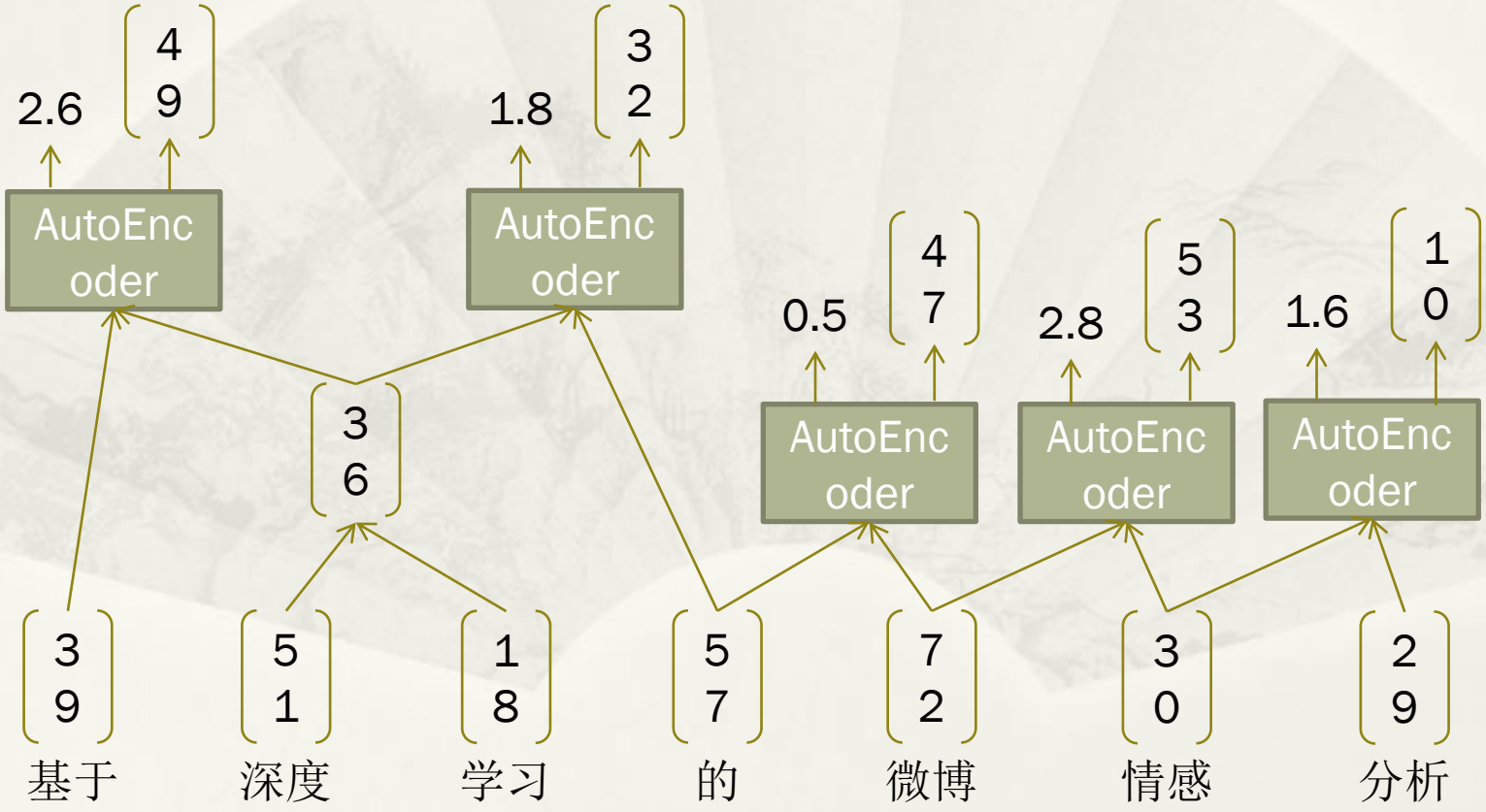
2.2 递归自编码(1/3)


IP
NP
VP
PU
CP
PP
VP
。
IP
DEC
NP
VV
VP
的
NN
NN
分析
VV
NP
微博
情感
基于
NN
NN
深度
学习

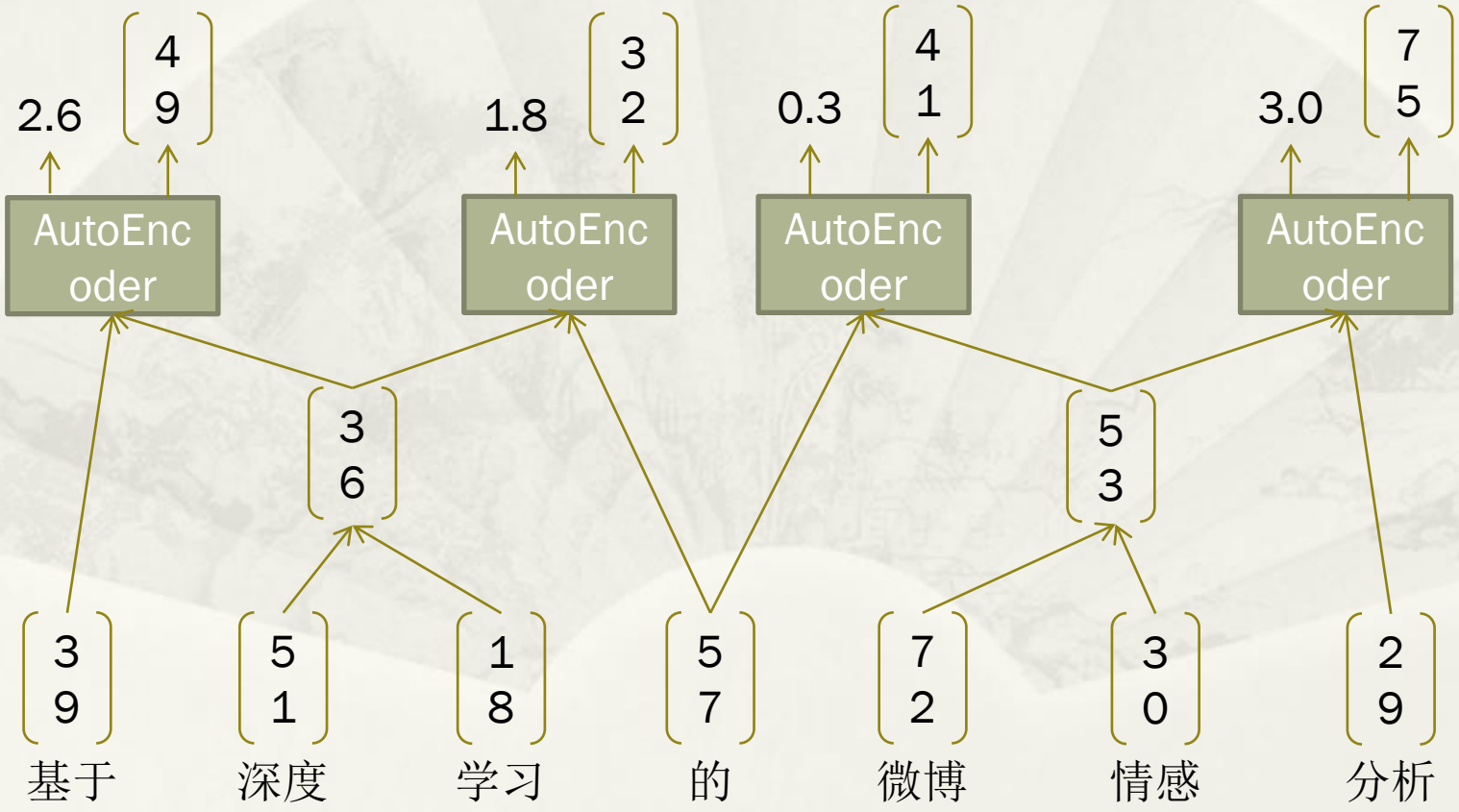
2.2 递归自编码(2/3)


0.4
1
9
3.4
3
6
0.9
3
3
0.5
4
7
2.8
5
3
1.6
1
0
AutoEncoder
AutoEncoder
AutoEncoder
AutoEncoder
AutoEncoder
AutoEncoder
3
9
5
1
1
8
5
7
7
2
3
0
2
9
基于
深度
学习
的
微博
情感
分析

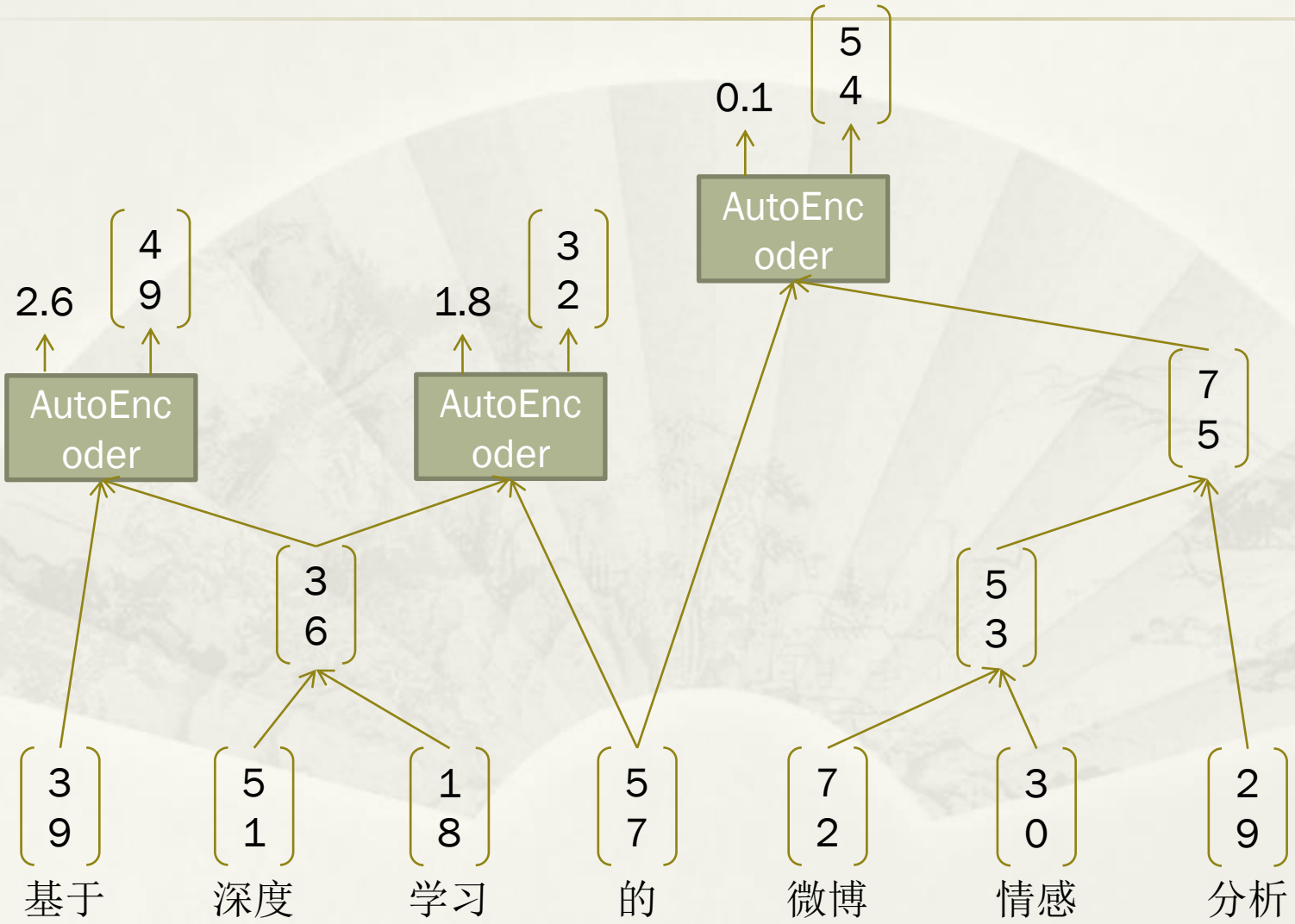
2.2 递归自编码(2/3)


2.6
4
9
AutoEnc
oder
1.8
3
2
AutoEnc
oder
3
6
0.5
4
7
AutoEnc
oder
2.8
5
3
AutoEnc
oder
1.6
1
0
AutoEnc
oder
3
9
基于
5
1
深度
1
8
学习
5
7
的
7
2
微博
3
0
情感
2
9
分析

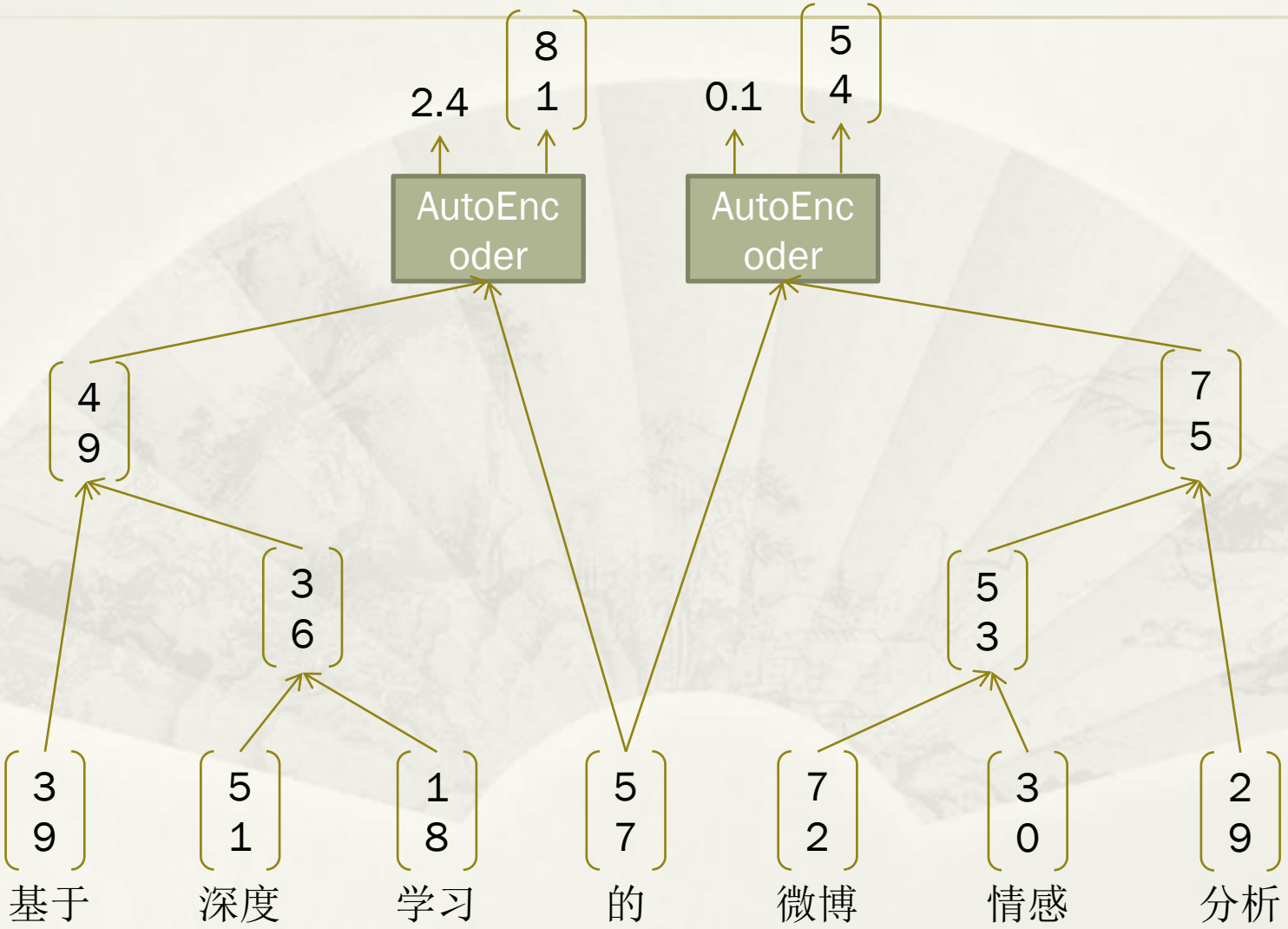
2.2 递归自编码(2/3)


2.6
4
9
1.8
3
2
0.3
4
1
3.0
7
5
AutoEncoder
AutoEncoder
AutoEncoder
AutoEncoder
3
6
5
3
3
9
5
1
1
8
5
7
7
2
3
0
2
9
基于
深度
学习
的
微博
情感
分析

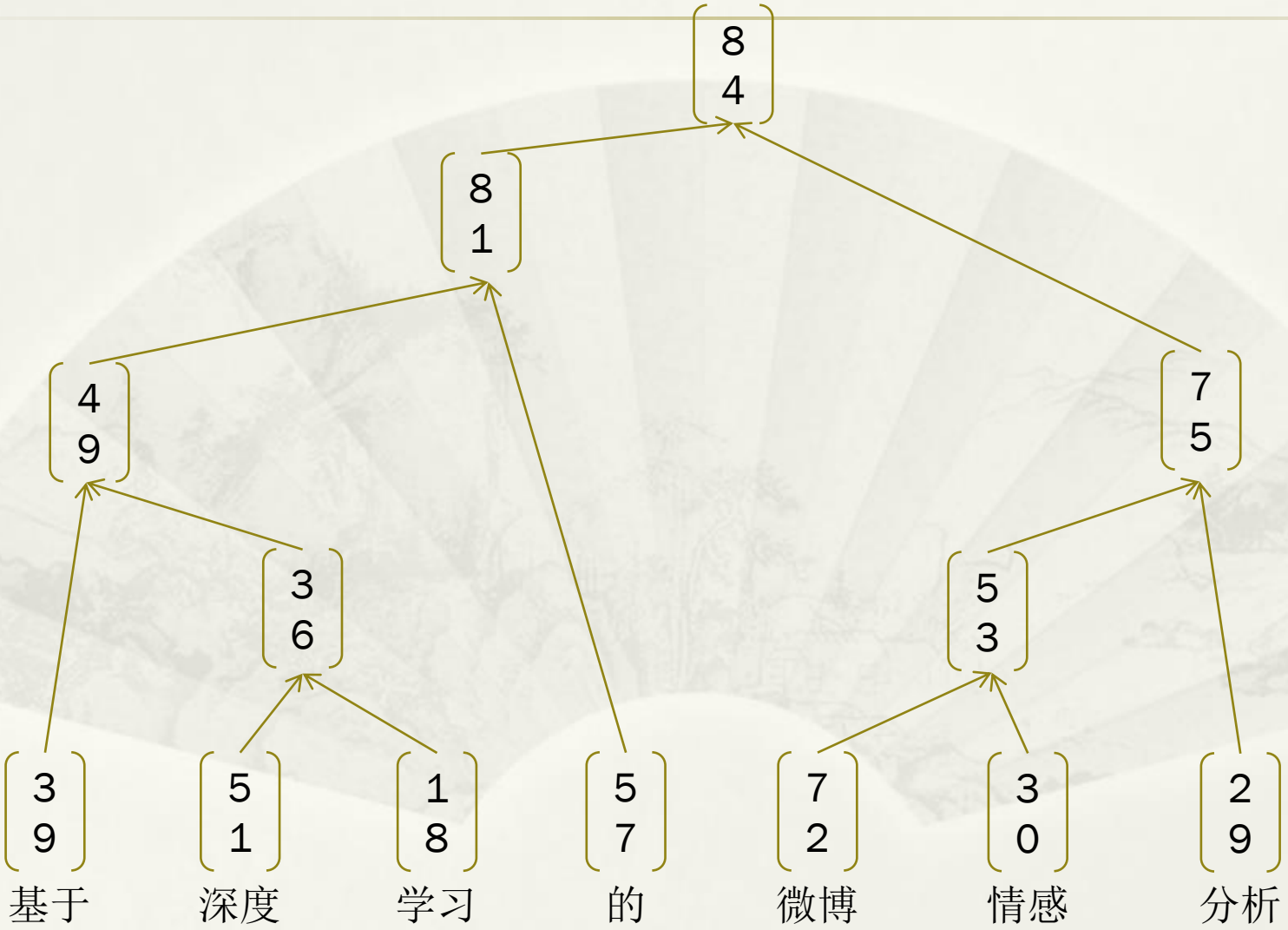
2.2 递归自编码(2/3)


基于
深度
学习
的
微博
情感
分析
AutoEncoder
AutoEncoder
AutoEncoder
2.6
1.8
0.1

2.2 递归自编码(2/3)

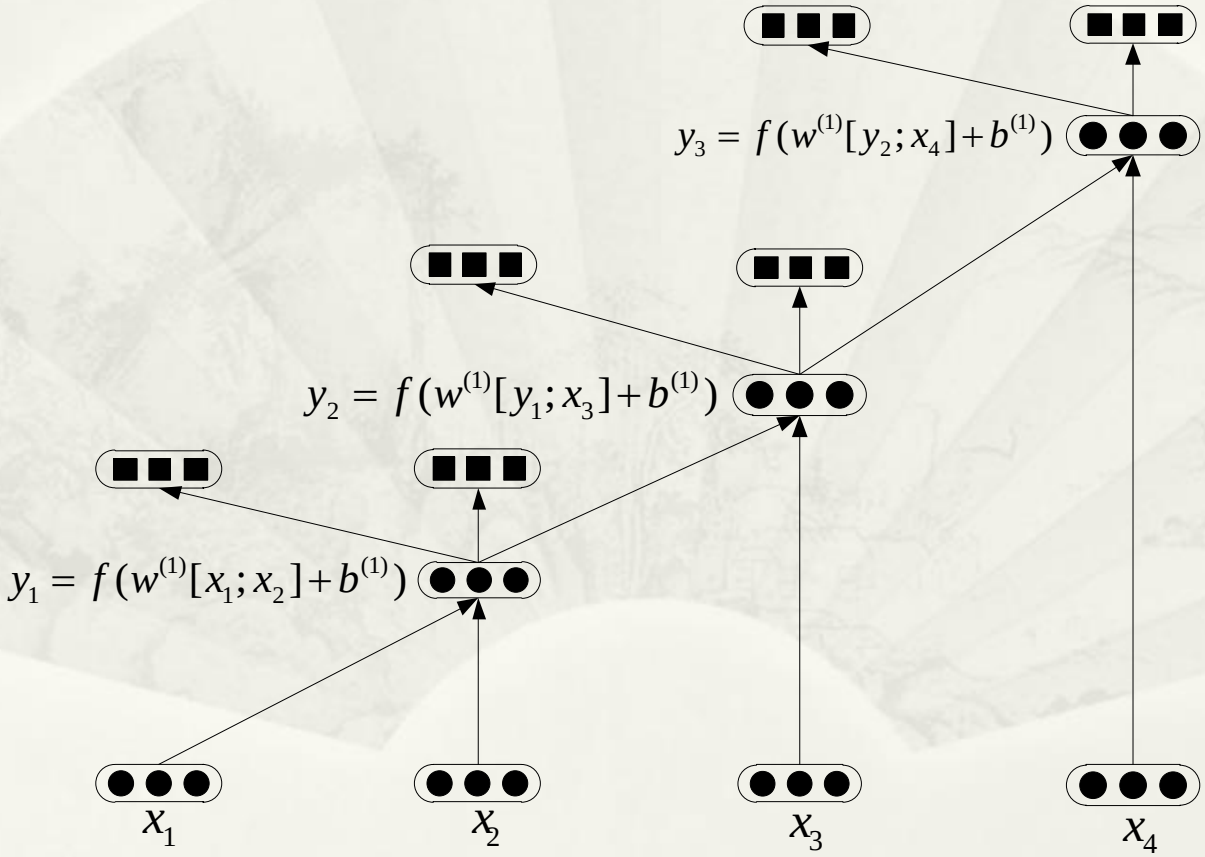

2.4
8
1
0.1
5
4
AutoEnc
oder
AutoEnc
oder
4
9
3
6
7
5
5
3
3
9
5
1
1
8
5
7
7
2
3
0
2
9
基于
深度
学习
的
微博
情感
分析

2.2 递归自编码(2/3)



2.2 递归自编码(3/3)

* 递归自编码（Recursive AutoEncoder）



报告提纲

1. 背景介绍
2. 微博情感分析模型
   - 2.1 词语的表示
   - 2.2 递归自编码(Recursive AutoEncoder)
   - 2.3 基于RAE的情感极性转移模型
   - 2.4 情感分析模型算法
3. 实验结果分析
4. 总结

2.3 基于RAE的情感极性转移模型 (1/3)

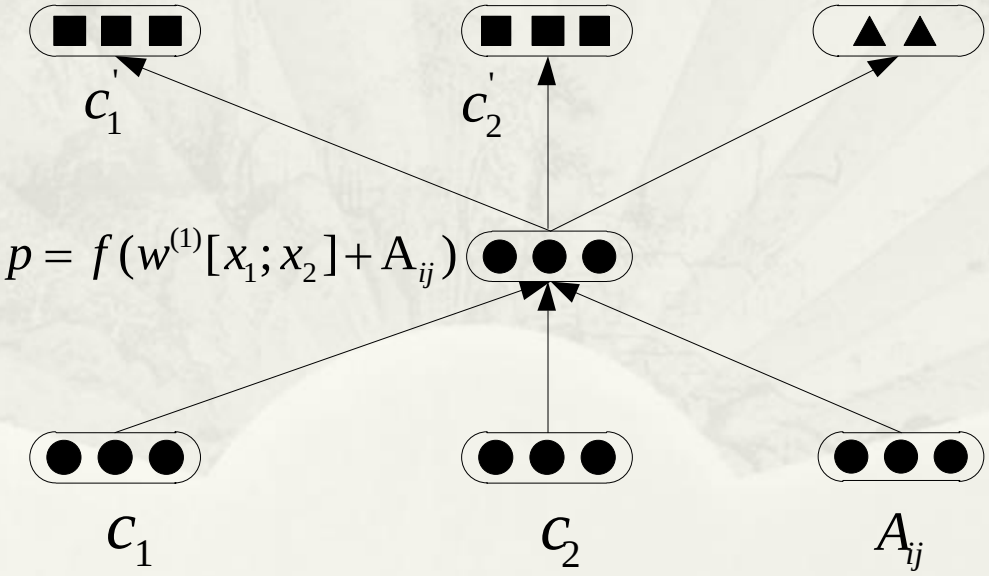
* 情感极性转移现象

极性转移类型实例	
正向+正向	她 很 漂亮 ！
正向+负向	他 很 自私 。
负向+正向	他 学习 不 好 。
负向+负向	他 学习 不 坏 。

2.3 基于RAE的情感极性转移模型(2/3)

- 情感极性转移模型

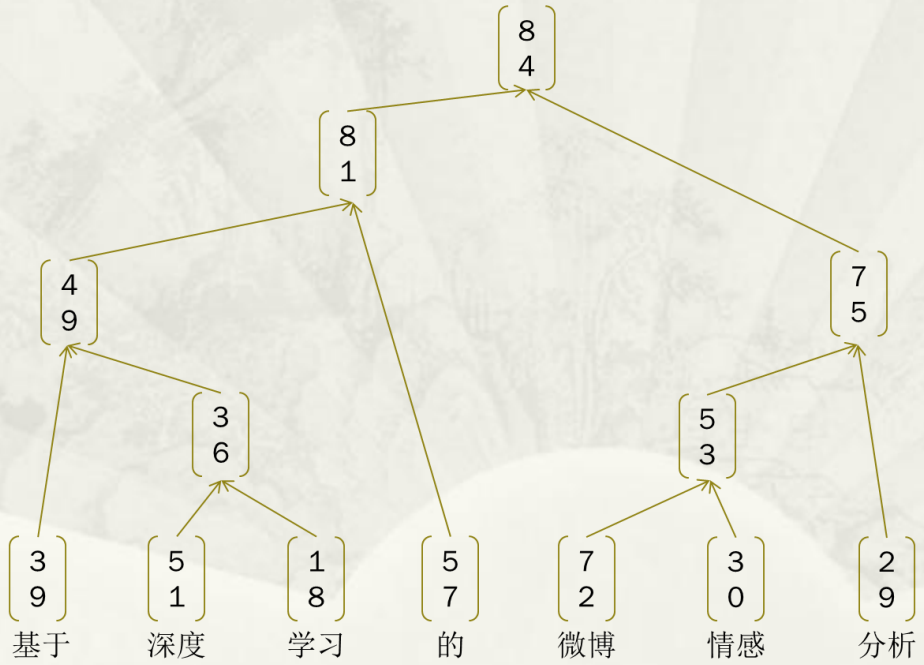
假定情感标注集T={正，负}，当连续两个词语的情感标签由T中的第i个t_i变为第j个t_j时，引入转换分数A_{ij}（$A_{ij} \in \{A_{01,}A_{00,}A_{11,}A_{10,}\}$）。



2.3 基于RAE的情感极性转移模型 (3/3)

- 模型输出

J=softmax(λ·x+b)，$\text{softmax} = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}$



报告提纲

1. 背景介绍
2. 微博情感分析模型
   - 2.1 词语的表示
   - 2.2 递归自编码(Recursive AutoEncoder)
   - 2.3 基于RAE的情感极性转移模型
   - 2.4 情感分析模型算法
3. 实验结果分析
4. 总结

2.4 情感分析模型算法

情感分析模型

输入：训练语料 S 及其对应标签 t

输出：$\theta = \{w^{(1)}, w^{(2)}, w^{(3)}, A, b^{(2)}, b^{(3)}\}$

1) 使用高斯分布初始化训练语料的词向量表示；

2) **while** 不收敛 **do**

$\nabla J = 0$

for all <s,t>ϵS **do**

使用贪心算法生成句子二叉树结构

计算 $\nabla J_i = \partial J(s,t)/\partial \Theta$

更新 $\nabla J = \nabla J + \nabla J_i$

end for

更新 $\Theta = \frac{1}{N}\nabla J + \lambda\theta$

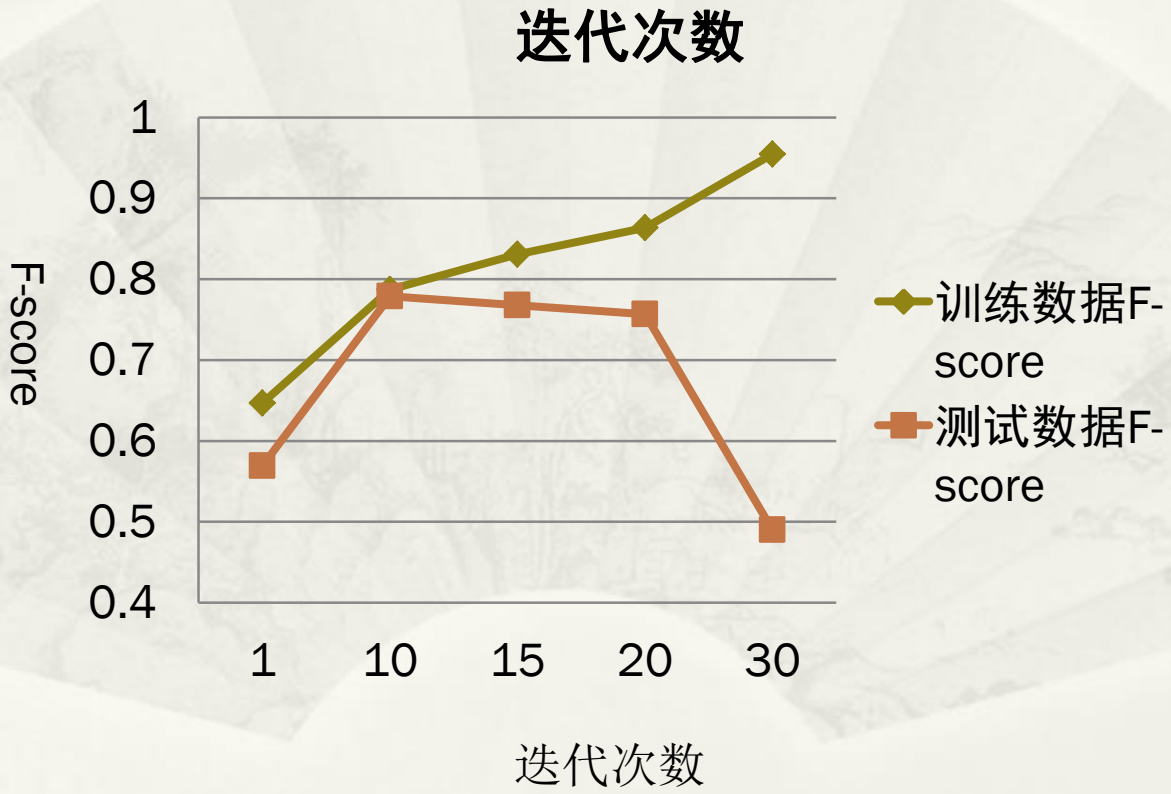
end while

报告提纲

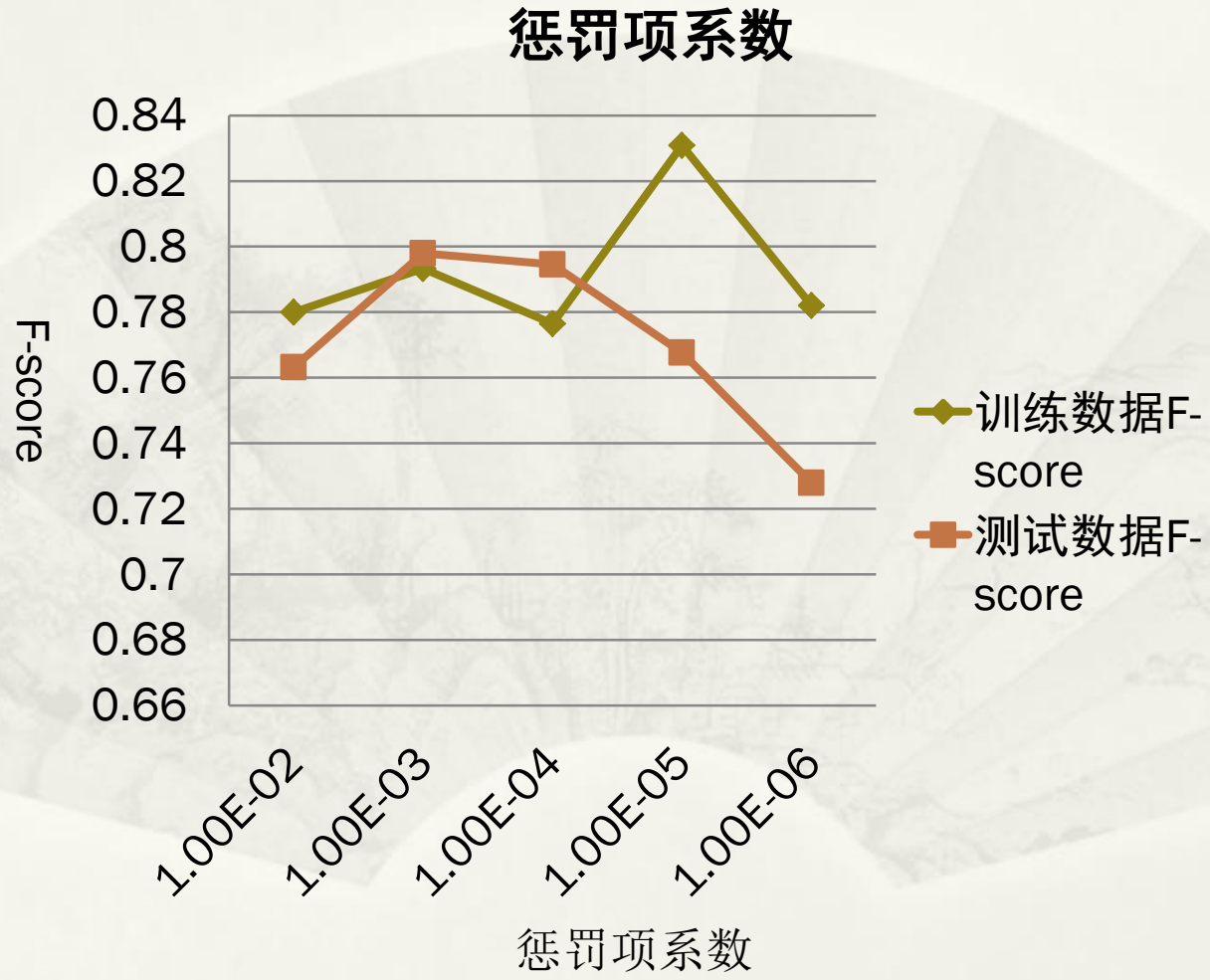
1. 背景介绍
2. 微博情感分析模型
   - 2.1 词语的表示
   - 2.2 递归自编码(Recursive AutoEncoder)
   - 2.3 基于RAE的情感极性转移模型
   - 2.4 情感分析模型算法
3. 实验结果分析
4. 总结

3. 实验结果分析(1/3)

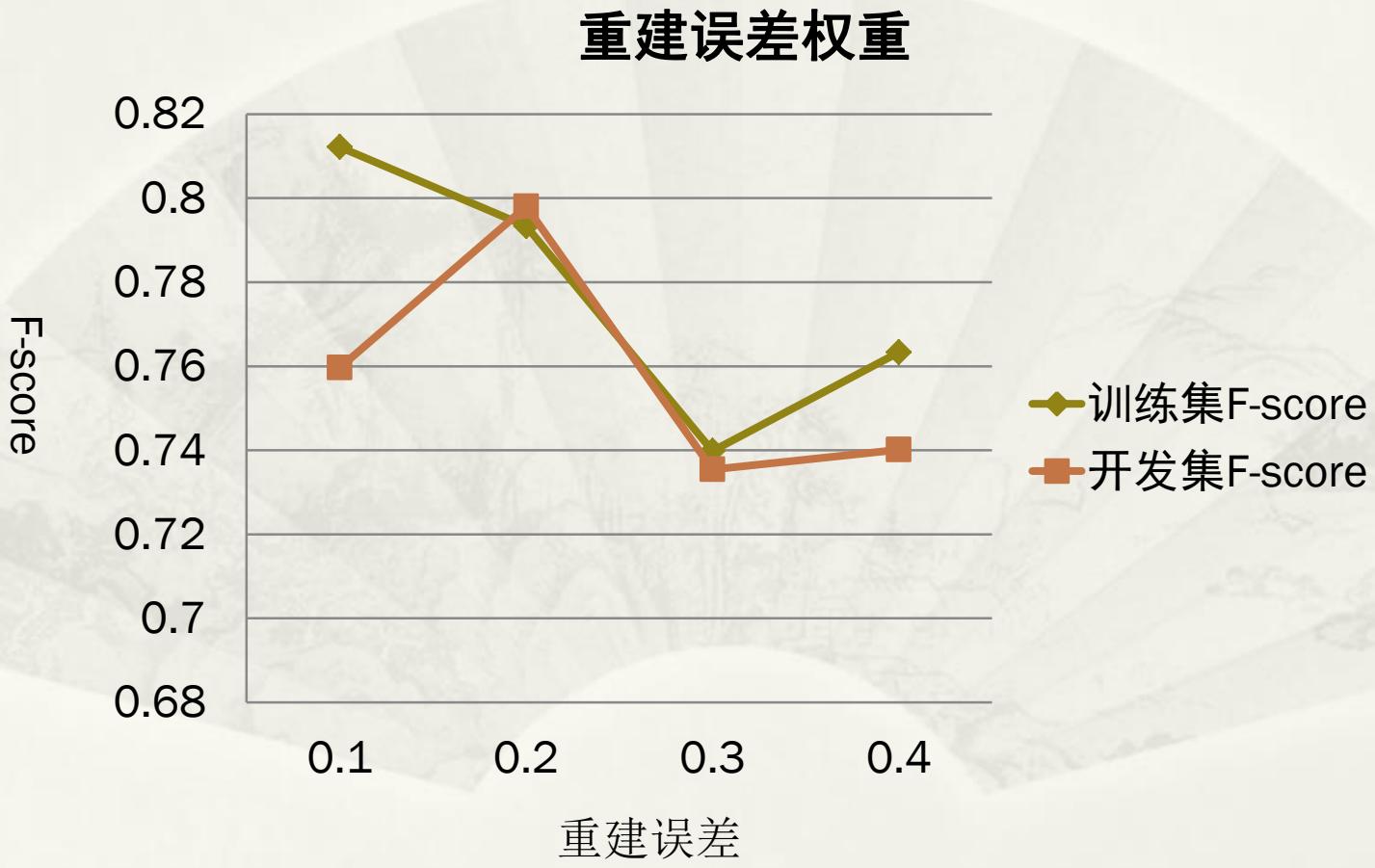
* 实验一：可调参数选择



3. 实验结果分析(1/3)

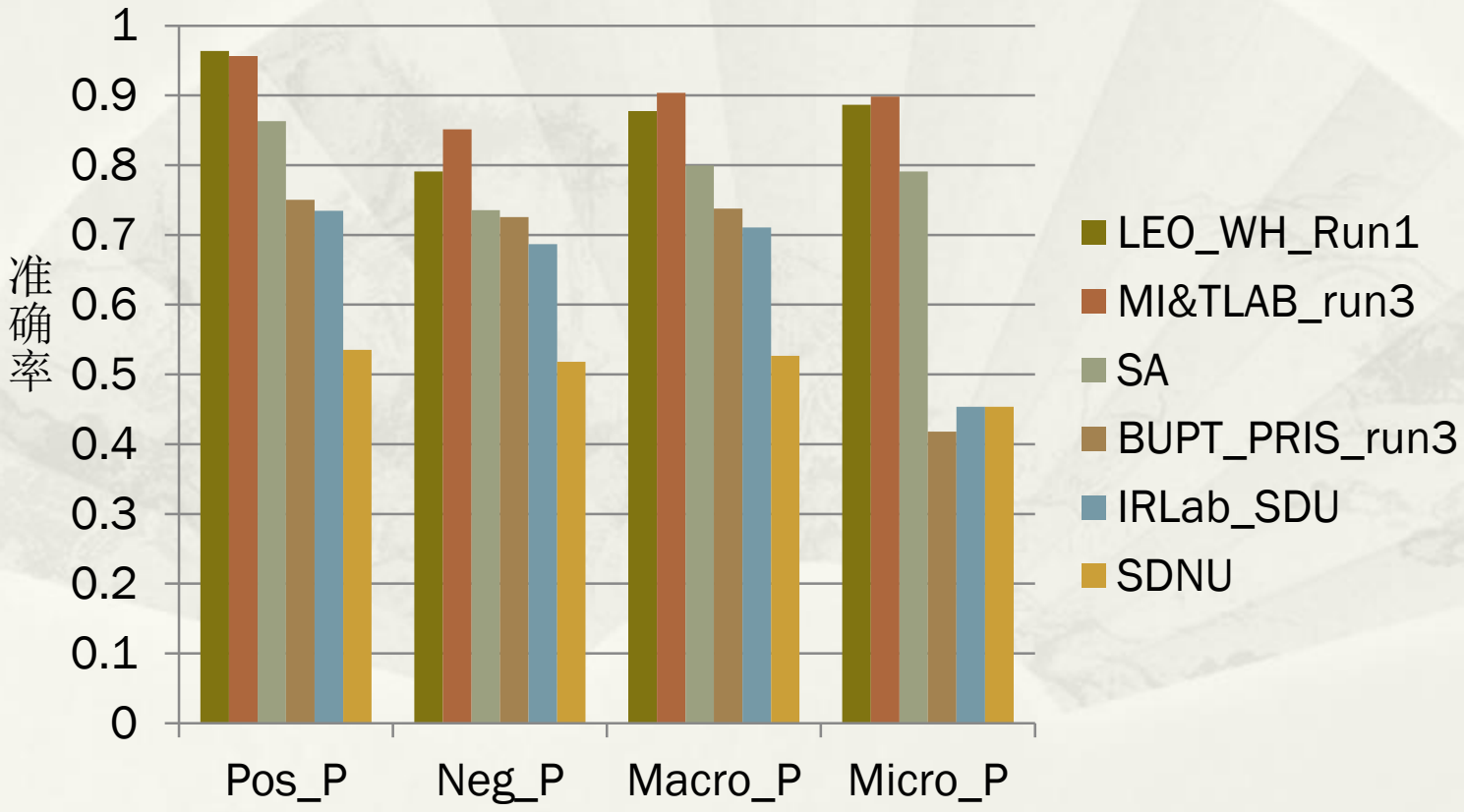

惩罚项系数
0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
F-score
1.00E-02
1.00E-03
1.00E-04
1.00E-05
1.00E-06
惩罚项系数
训练数据F-score
测试数据F-score

3. 实验结果分析(1/3)


重建误差权重
0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.68
F-score
0.1
0.2
0.3
0.4
重建误差
训练集F-score
开发集F-score

3. 实验结果分析(2/3)

* 实验二：COAE2014数据集封闭测试



3. 实验结果分析(3/3)

* 实验三：对比试验

参数设定

可调参数	值
词向量维数	50
迭代次数	10
正则惩罚项系数	1e-03
重建误差与交叉熵误差权重比	0.2

3. 实验结果分析(3/3)

模型结果对比

系统	准确率(%)	召回率(%)	F-Score(%)
RAE	78.75	78.44	77.88
情感极性转移模型	80.59	80.59	79.80

报告提纲

1. 背景介绍
2. 微博情感分析模型
   - 2.1 词语的表示
   - 2.2 递归自编码(Recursive AutoEncoder)
   - 2.3 基于RAE的情感极性转移模型
   - 2.4 情感分析模型算法
3. 实验结果分析
4. 总结

4. 总结

- 本文将深度学习方法应用到中文微博情感分析达到与使用手工标注特征相当的水平。
- 根据语言语法现象改进现有模型提出情感极性转移模型，较原模型效果有一定提升。

4. 进一步工作(1/2)

* 中文中的语法现象

否定：机器的智能不会超过人脑。

加强：深度学习的应用会更加广泛。

比较：小明比小红的个子高。

* 中文中的多义现象

例如：先帝不以臣卑鄙。

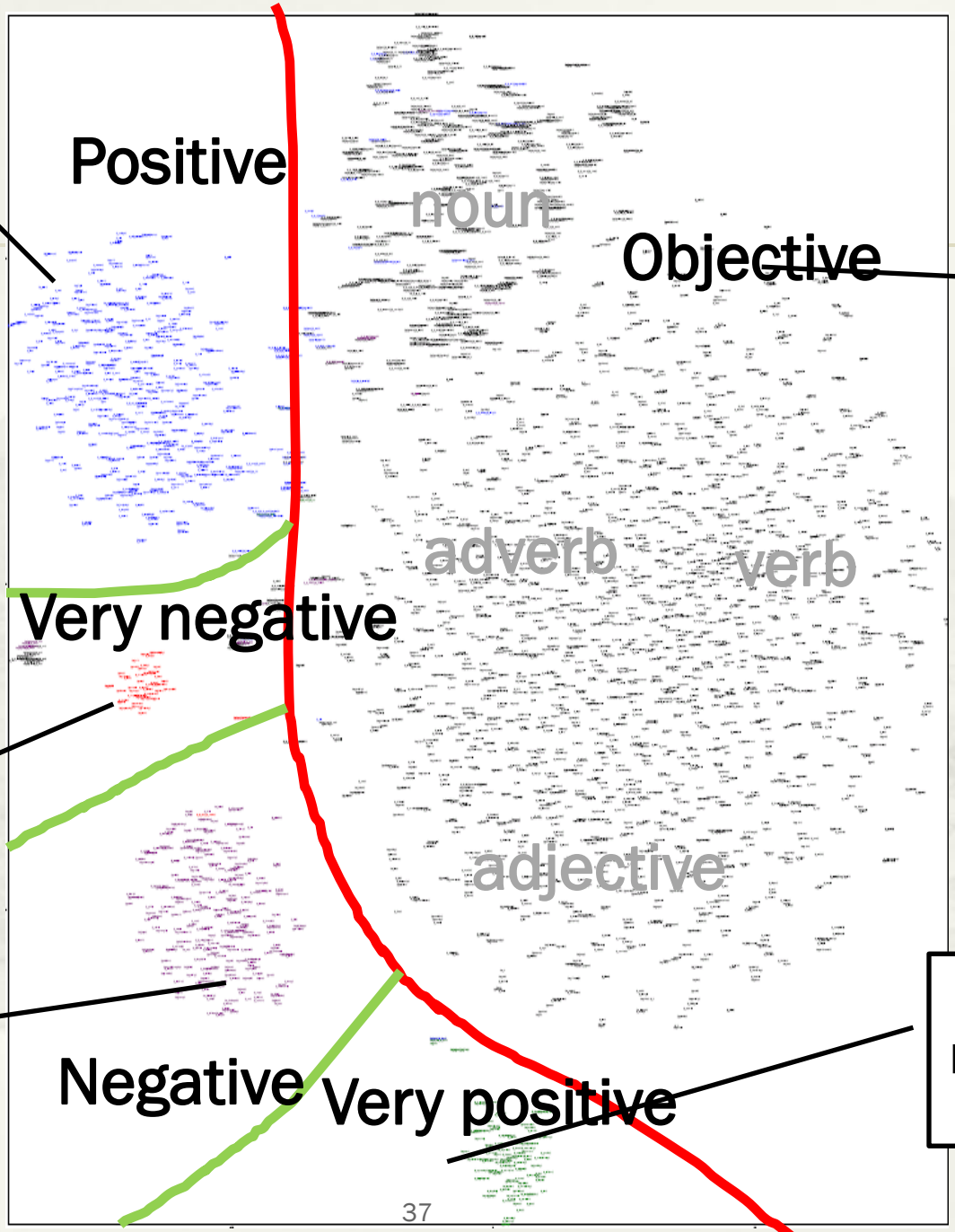
——摘自　诸葛亮《前出师表》

一些势利小人卑鄙，惯会兴风作浪，挑拨离间。

——摘自　茅盾《霜叶红似二月花》

4. 进一步工作(2/2)

* 深度学习在聚类上的应用



fancy, good, cool, promising, interested

plot, near, bu surface, ther version

failure, worst, disaster, horrible

problem, slow, sick, mess, poor, wrong

creative, great, perfect, superb, amazing

敬请各位老师同学指正！
谢谢！

梁军
邮箱：*zhengdaxg@163.com*
郑州大学自然语言处理实验室
实验室网站：*http://nlp.zzu.edu.cn*