



情感分析及传统求解方法

(Sentiment Analysis and Its Traditional Solutions)

刘远超

哈尔滨工业大学

计算机科学与技术学院

什么是情感分析？

情感分析： 又称意见挖掘、倾向性分析等。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。



Chris蕊 Lv4 VIP

★★★★★ 口味：非常好 环境：非常好 服务：非常好 人均：60元

新开的一家喝精酿啤酒吃下酒小菜的地方，位置很好找，店不是很大，一共也就20张桌吧...可以说环境非常好，超大的投影，看起来很棒。

菜品 有牛排，品质很好，值得称赞的是很多下酒小菜是你在外面吃不到的，那款烤鹅蛋就很惊艳，在外面绝对吃不到，还有烤猪心，特别适合喝酒。

啤酒 更加惊艳，20多杯酒一起上来 由2个服务员拖过来 精酿的啤酒品质非常棒...

情感分析的常见研究内容：

- 判断一段文本是positive 还是 negative?
- 预测一段文本的情感分值：1-5
- 抽取评论文本中的评价对象，以及与其相对应的情感倾向。

情感分类和文本分类的关系

- 文本分类（Text classification）：
 - 根据文档的主题信息分类
 - 主题词很重要
- 情感分类（Sentiment classification）：
 - 根据文档的情感信息分类
 - 情感词很重要
 - 与文档分类类似但有一定的不同

现有情感词典

- 英文情感词典举例

- <http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>

- 2006 positive,4783 negative

```
positive-words.txt
40 abundant
41 accessible
42 accessible
43 acclaim
44 acclaimed
45 acclamation
46 accolade
47 accolades
48 accommodative
49 accomodative
50 accomplish
51 accomplished
52 accomplishment
```

```
negative-words.txt
40 abominable
41 abominably
42 abominate
43 abomination
44 abort
45 aborted
46 abortions
47 abrade
48 abrasive
49 abrupt
50 abruptly
51 abscond
52 absence
```

- 中文情感词典举例

- HowNet第一版（褒贬词）

```
2187
2188 NO.=000244
2189 W_C=蔼然可亲
2190 G_C=ADJ
2191 E_C=
2192 W_E=affably
2193 G_E=ADV
2194 E_E=
2195 DEF=aValue|属性值,behavior|举止,kindhearted|善,desired|良
2196
7569
7570 NO.=000842
7571 W_C=暗无天日
7572 G_C=ADJ
7573 E_C=
7574 W_E=in complete darkness
7575 G_E=PP
7576 E_E=
7577 DEF=aValue|属性值,SocialMode|风气,bad|坏,undesired|莠
7578
```

情感词典构建方法之一

- 基于现有情感词典和通用词典（如**wordnet**）的扩充方法

- WordNet: Princeton 大学开发的一种基于认知语言学的[英语词典](https://wordnet.princeton.edu/)。

- <https://wordnet.princeton.edu/>

- 步骤:

1. 创建正向的种子词(如“good”)和负向的种子词(如“terrible”);
2. 在通用词典中找到这些词的同义词和反义词, 更新情感词典;
 - I. **Positive Set:** 将正向词的同义词(如“well”), 以及反向词的反义词保存;
 - II. **Negative Set:** 将反向词的同义词(如“awful”), 以及正向词的反义词;

Kim, Soo Min. "Determining the sentiment of opinions." *International Conference on Computational Linguistics Association for Computational Linguistics*, 2004:1367.

情感词典构建方法之二

- 基于现有情感词典及网络生语料的情感词典扩充方法
- 直觉：
 - 用 “*and*” 连接的形容词一般具有相同情感倾向
 - Fair **and** legitimate, corrupt **and** brutal
 - 用 “*but*” 等转折词连接的形容词一般具有相反的情感倾向
 - fair **but** brutal

情感词典构建方法二（续）

- 基于现有情感词典及网络生语料的情感词典扩充方法
 - Step 1: 人工获得若干种子情感词的集合，如种子词“nice”；
 - Step 2: 对种子情感词在生语料上查询扩充，得到新的情感词。

The beach **was nice and** clean although the ... 翻译此页

位置: Playas Uvero Alto, Punta Cana

Excellence Punta Cana: The beach **was nice and** clean although the water... - See 13,702 traveler reviews, 15,063 candid photos, **and** great deals for Excellence Punta Cana at ...

<https://www.tripadvisor.com/ShowUserReviews-g147293-d218524-r...> ▼ 2009-9-23

Accommodation in London - ... 翻译此页

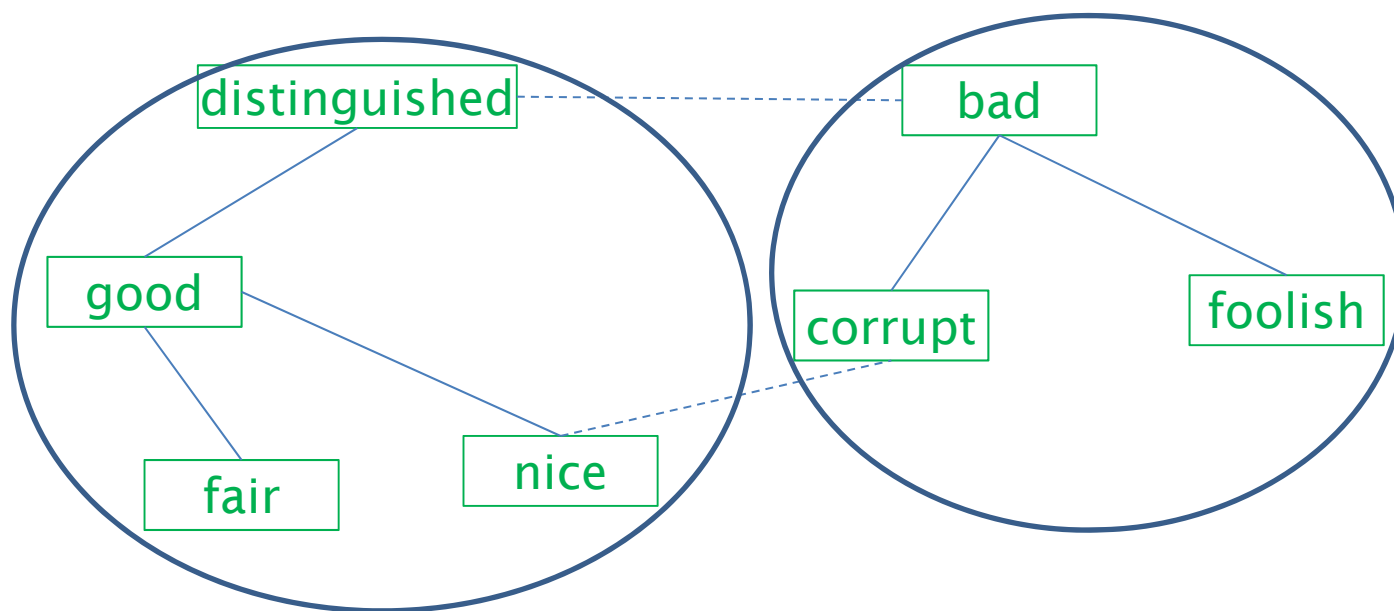
The apartment **was nice and** clean as it had all the equipment needed . I came very satisfied with your service as my son too. Thanks - Hazel, Portugal. ...

nikeapartments.com ▼

Hatzivassiloglou V, Mckeown K R. Predicting the Semantic Orientation of Adjectives[J]. Proceedings of the ACL, 1997:174--181.

情感词典构建方法二（续）

- Step 3: 根据每个词对之间的极性相似度（polarity similarity），得到一个图；
- Step 4: 采用聚类的方法，分割为两个类（positive和negative）。



典型数据集以及传统分类步骤

- 典型数据集:

- 影评数据集: *Polarity Data 2.0*:

- ◆ <http://www.cs.cornell.edu/people/pabo/movie-review-data>

- ◆ Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. [*Thumbs up? Sentiment Classification using Machine Learning Techniques*](#). *EMNLP-2002*, 79—86.

- 传统分类步骤:

- 词语切分 (Tokenization)

- 特征抽取 (Feature Extraction) 以及向量化表示等。

- 分类器

- ✓ Naïve Bayes

- ✓ SVM

如何处理否定现象

- 例如:

- I **didn't** like this movie

- I **really** like this movie

- 一种简单的办法: 在否定词和之后最近的标点符号之间的每个词前面加上 “**NOT_**”:

didn't like this movie , but I

didn't NOT_like NOT_this NOT_movie, but I.....

1. Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*.
2. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *EMNLP-2002*, 79—86.

情感分类的难点

- 举例:

- “这款手机用了不几天就没电了”
- “汽车噪音太大”
- “手机屏幕很大”
- “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”----香水的评论
- “ not very good ”

Thanks!





词向量

(Word Vector)

刘远超

哈尔滨工业大学

计算机科学与技术学院

传统的语义向量表示方法

- 自然语言处理中**独热编码下词的表示**：每个词表示为一个很长的向量。向量的维度是词表大小，其中只有一个维度的值为1，其他元素为0。
例如：
“菠萝”表示为 [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 ...]
“凤梨”表示为 [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 ...]
- **文档的表示**一般采用词袋BOW模型：如果A文档含有“菠萝”，B文档含有“凤梨”，则由于二者分别在不同的维度上，仍然很难看出关联。
- 可见传统表示方法存在的一个问题就是“词汇鸿沟”现象。

深度学习中的词向量表示

- 比较而言，深度学习中的词向量通常使用一种低维实数向量，一般称之为Distributed representation、word embedding等。
 - 例如：[0.122, -0.632, -0.932, -0.739, -0.92, ...]。
 - 维度一般为 50 维、100 维、200维或者300维等。
- 这使得在语义上相似或相关的词，在距离上更接近。
 - 例如，此时“菠萝”和“凤梨”的相似度会较大。

利用Word2vec得到词向量

- Word2vec 是 Google公司在 2013 年开源的将词表示为低维实数值向量的工具。
- 具体步骤：
 - ① 从<http://word2vec.googlecode.com/svn/trunk/> 下载代码并解压缩；
 - ② 下载训练例子语料文件：从<http://mattmahoney.net/dc/text8.zip> (解压后不到100M，可解压放到与word2vec的同级目录下)
 - ③ 运行make编译word2vec工具：
 - ④ 对训练语料进行训练。如输入命令：
`# ./word2vec -train text8 -output vectors.bin -cbow 0 -size 48 -window 5 -negative 0 -hs 1 -sample 1e-4 -threads 20 -binary 0 -iter 100`
#binary 设为1表示结果用二进制存储，为0是普通存储（可以看到词语和对应的向量）
 - ⑤ 向量的输出文件为vectors.bin。

利用word2vec进行词的相似度计算

```
#./distance vectors.bin
```

```
Enter word or sentence (EXIT to break): china
```

```
Word: china Position in vocabulary: 486
```

```
Word Cosine distance
```

```
-----
```

```
taiwan 0.656181
```

```
japan 0.633499
```

```
tibet 0.607813
```

```
manchuria 0.581230
```

```
hainan 0.561931
```

```
xiamen 0.555860
```

```
chongqing 0.550099
```

```
jiang 0.549195
```

```
chinese 0.548320
```

GloVe: Global Vectors for Word Representation

- 项目网址: <https://nlp.stanford.edu/projects/glove/>
- 代码: <https://github.com/stanfordnlp/GloVe>
- 论文: Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [GloVe: Global Vectors for Word Representation](#). EMNLP 2014. 1532-1543

Thanks!





递归神经网络及其变体

(Recursive Neural Networks and Its Variants)

刘远超

哈尔滨工业大学

计算机科学与技术学院

如何判断一段文本的情感倾向？

以“**not very good**”为例：

- 深度学习方法：

- 词义如何表示：每个词表示为一个低维向量

维度	1	2	3	4	5	6	7	8	9	10
not:	-0.022975	0.087888	-0.24248	-0.074691	0.02824	0.22998	-4.5327	0.5373	-0.12518	-0.66138
very:	0.24487	-0.15216	-0.3001	-0.22925	0.071235	-0.37077	-4.2901	0.068465	-0.31345	-0.7891
good:	-0.069254	0.37668	-0.16958	-0.27482	0.25667	-0.20293	-4.1122	0.02595	-0.27085	-0.87003

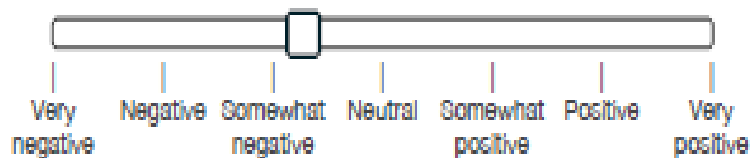
- 问题的关键：语义的组合和分类
- Socher 等人采用情感树库和神经网络模型进行处理：
 - 单个语句的两类情感分类精度从80% 提高到 85.4%
 - 对于短语级别的情感预测精度达到了从71%提高到80.7%。

Socher et al.. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions, EMNLP 2011

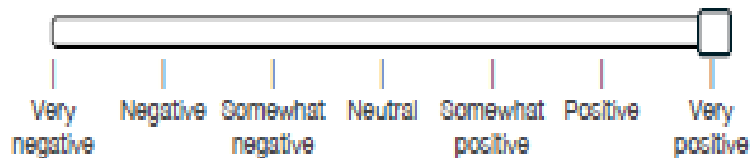
情感树库语料

- 情感树库（**Stanford Sentiment Treebank**）中的数据来源于Pang and Lee(2005)发布的电影评论（roottentomatoes.com）语料。
- 标注过程：用Stanford Parser 进行句法分析，然后在此基础上进行人工标注。图中的滑块有25个不同的值，其初始值设为neutral。

nerdy folks



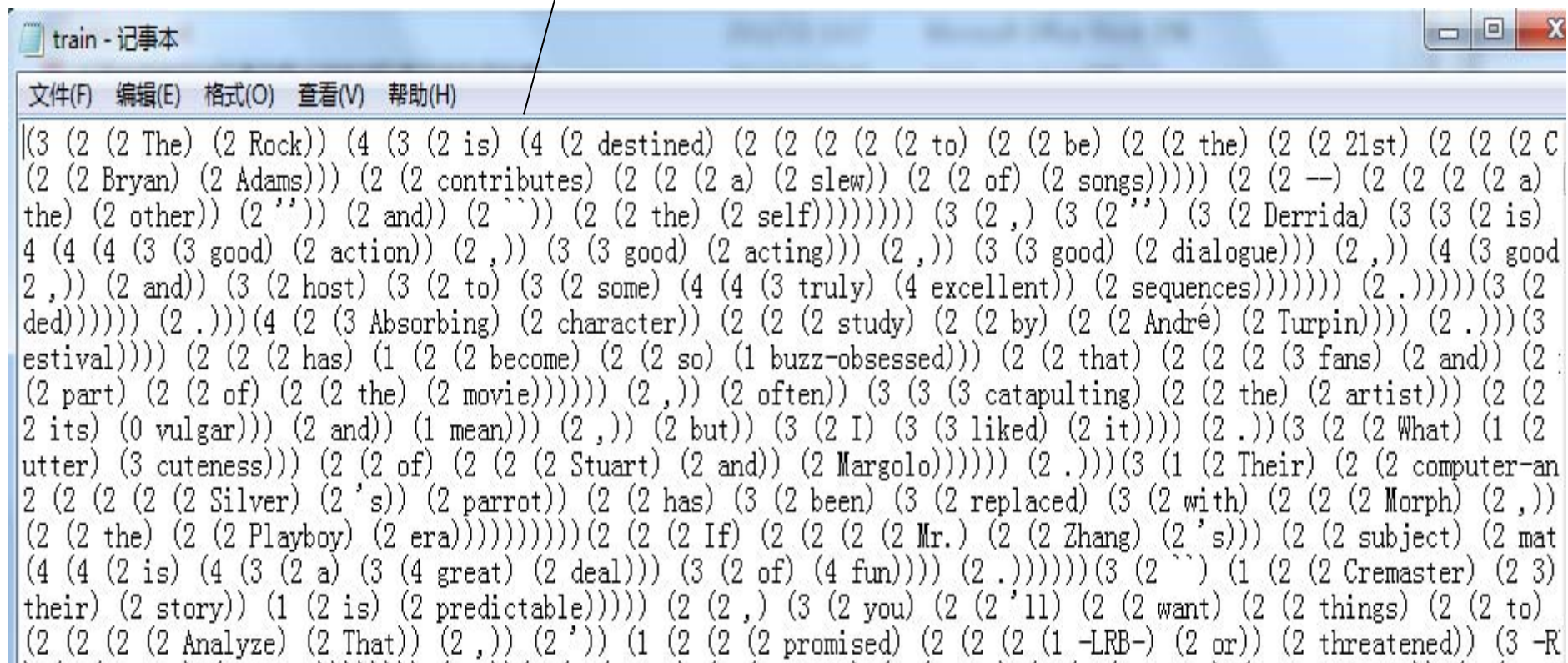
phenomenal fantasy best sellers



- 语料规模：对11855个句子中的215,154个短语（**phrases**）进行了标注。

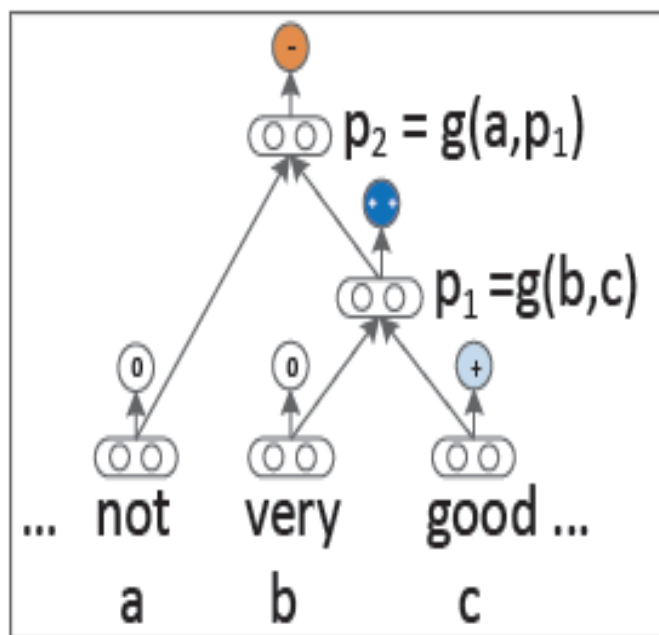
数据描述（续）

5类：0-4



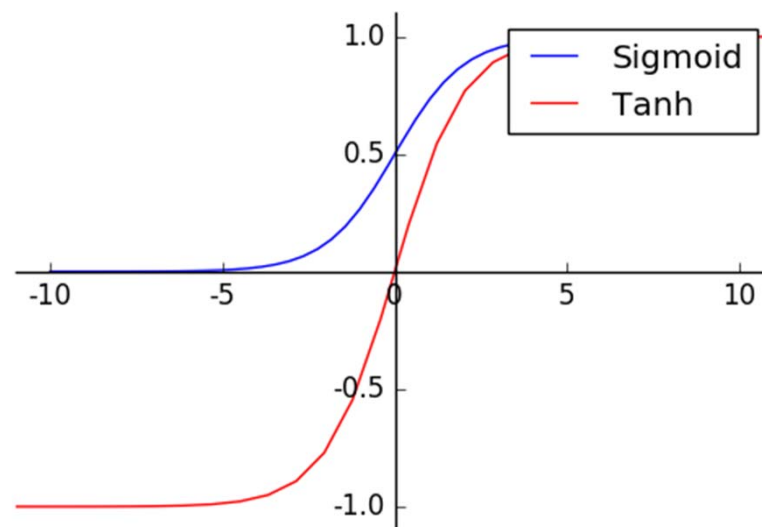
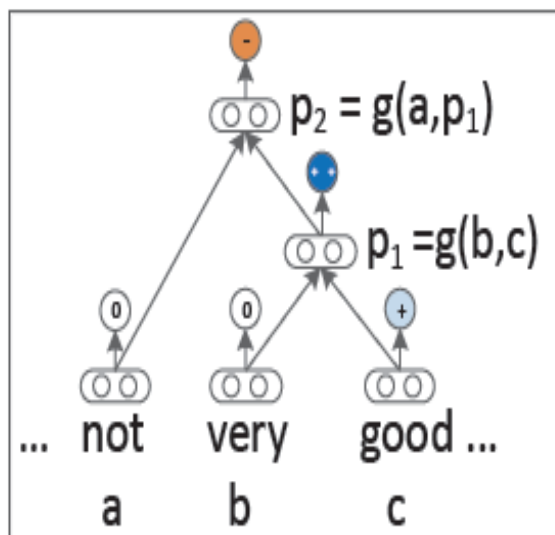
```
train - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
| (3 (2 (2 The) (2 Rock)) (4 (3 (2 is) (4 (2 destined) (2 (2 (2 (2 to) (2 (2 be) (2 (2 the) (2 (2 21st) (2 (2 (2 C
(2 (2 Bryan) (2 Adams))) (2 (2 contributes) (2 (2 (2 a) (2 slew)) (2 (2 of) (2 songs)))))) (2 (2 --) (2 (2 (2 a)
the) (2 other)) (2 ')) (2 and)) (2 ``)) (2 (2 the) (2 self)))))) (3 (2 ,) (3 (2 ')) (3 (2 Derrida) (3 (3 (2 is)
4 (4 (4 (3 (3 good) (2 action)) (2 ,)) (3 (3 good) (2 acting)) (2 ,)) (3 (3 good) (2 dialogue)) (2 ,)) (4 (3 good
2 ,)) (2 and)) (3 (2 host) (3 (2 to) (3 (2 some) (4 (4 (3 truly) (4 excellent)) (2 sequences)))))) (2 .)) (3 (2
ded)))))) (2 .)) (4 (2 (3 Absorbing) (2 character)) (2 (2 (2 study) (2 (2 by) (2 (2 André) (2 Turpin))) (2 .)) (3
estival)))) (2 (2 (2 has) (1 (2 (2 become) (2 (2 so) (1 buzz-obsessed)) (2 (2 that) (2 (2 (2 (3 fans) (2 and)) (2
(2 part) (2 (2 of) (2 (2 the) (2 movie)))))) (2 ,)) (2 often)) (3 (3 (3 catapulting) (2 (2 the) (2 artist)) (2 (2
2 its) (0 vulgar))) (2 and)) (1 mean)) (2 ,)) (2 but)) (3 (2 I) (3 (3 liked) (2 it))) (2 .)) (3 (2 (2 What) (1 (2
utter) (3 cuteness))) (2 (2 of) (2 (2 (2 Stuart) (2 and)) (2 Margolo)))) (2 .)) (3 (1 (2 Their) (2 (2 computer-an
2 (2 (2 (2 Silver) (2 's)) (2 parrot)) (2 (2 has) (3 (2 been) (3 (2 replaced) (3 (2 with) (2 (2 (2 Morph) (2 ,))
(2 (2 the) (2 (2 Playboy) (2 era)))))) (2 (2 (2 If) (2 (2 (2 Mr.) (2 (2 Zhang) (2 's)) (2 (2 subject) (2 mat
(4 (4 (2 is) (4 (3 (2 a) (3 (4 great) (2 deal))) (3 (2 of) (4 fun))) (2 .)) (3 (2 ``) (1 (2 (2 Cremaster) (2 3)
their) (2 story)) (1 (2 is) (2 predictable)))) (2 (2 ,) (3 (2 you) (2 (2 'll) (2 (2 want) (2 (2 things) (2 (2 to)
(2 (2 (2 (2 Analyze) (2 That)) (2 ,)) (2 ')) (1 (2 (2 (2 promised) (2 (2 (2 (1 -LRB- (2 or)) (2 threatened)) (3 -R
```

标准RNN (Socher et al., 2011)



- 每个词初始化表示为一个 d 维的向量。由一个随机均匀分布(uniform distribution) $U(-r; r)$, $r = 0.0001$ 随机采样生成
- 所有的词向量被存储在一个词嵌入矩阵 $L \in R^{d \times |V|}$ 中。其中 $|V|$ 是词汇表的大小
- 父节点的向量如何获得？利用组合函数 g 。

标准RNN（续）



- 父向量的计算方法:

- 以 p_1 为例, p_1 节点对应的的向量的计算方法为

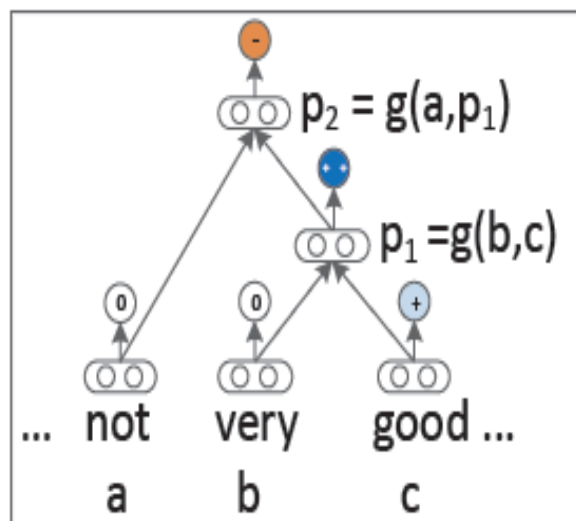
$$\mathbf{p}_1 = f(W \begin{bmatrix} b \\ c \end{bmatrix}),$$

其中, $W \in R^{d \times 2d}$

而 $f = \tanh$, $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ 。

标准RNN（续）

●情感分类概率分布的计算：



- 假设节点p1的向量为 a ，将其传给softmax分类器，从而计算情感分类概率分布 y^a ：

$$y^a = \text{softmax}(W_s * a)$$

$$\text{softmax}_i = \frac{\exp(x_i)}{\sum_{j=1}^c \exp(x_j)}$$

其中， $W_s \in R^{5 \times d}$ ， $a \in R^{d \times 1}$ ，则 $W_s * a \in R^{5 \times 1}$ 。

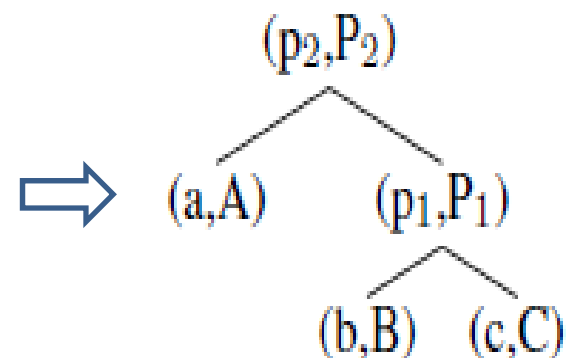
●不同模型的区别在于如何采用自底向上的方法计算隐含向量！

MV-RNN

- MV-RNN(Socher et al., 2012)的基本思想是将每个词或者父节点表示为一个向量和一个矩阵。
- 每个词的矩阵被初始化为一个 $d \times d$ 的单位矩阵identity matrix，然后再加上少量的高斯噪音。
- 父节点的向量和矩阵的计算方法分别为

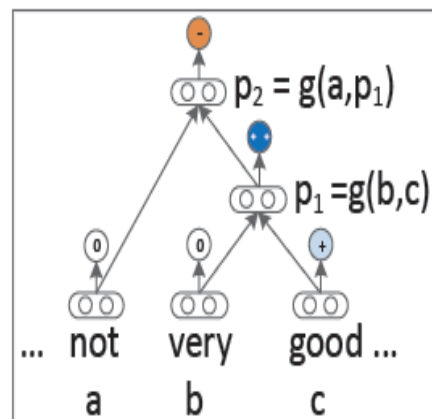
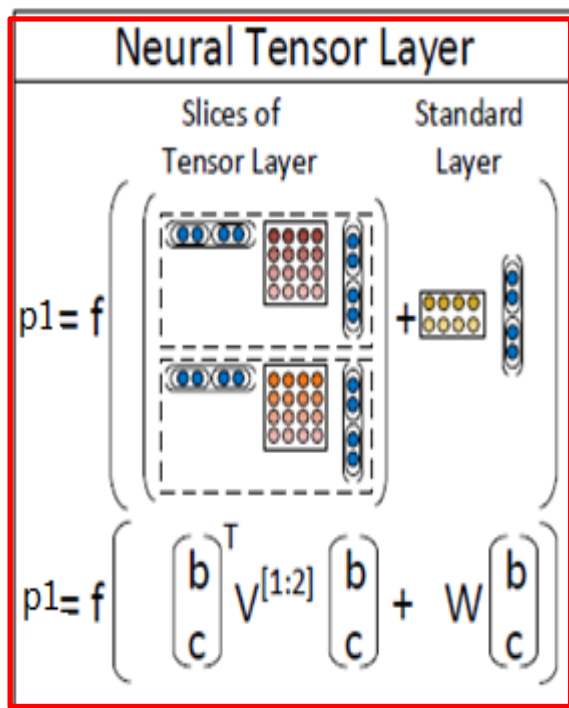
$\mathbf{p}_1 = f(W[\begin{smallmatrix} \mathbf{c} \\ \mathbf{b} \end{smallmatrix}]))$ ，其中， $W \in R^{d \times 2d}$ ，结果 p_1 是 $d \times 1$ 的向量。

$\mathbf{P}_1 = f(W_M[\begin{smallmatrix} \mathbf{B} \\ \mathbf{C} \end{smallmatrix}]))$ ，其中， $W_M \in R^{d \times 2d}$ ，结果 P_1 是 $d \times d$ 的矩阵。



Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]// Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012:1201-1211.

RNTN



- Recursive Neural Tensor Network
- 引入了张量层 (tensor layer), 包含d个slice。
- $V^{[1:d]} \in R^{2d \times 2d \times d}$, $W \in R^{d \times 2d}$

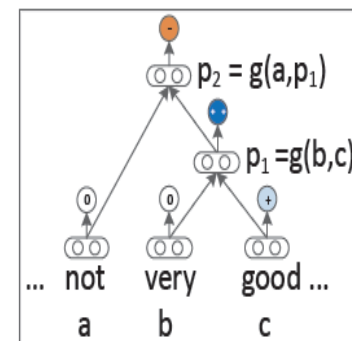
思考:

- 1) 该结构有什么优点? 和标准RNN的关系?
- 2) 图中的两个红色矩阵是否相同?
- 3) 红色矩阵为什么是 $2d \times 2d$ 维的, 且要有d个这样的矩阵?

RNTN的交叉熵损失函数

- 计算在节点 i 上的预测情感类分布 $y^i \in R^{C \times 1}$ 与真实情感类分布 $t^i \in R^{C \times 1}$ 之间的交叉熵损失:

$$E(\theta) = -\sum_i \sum_j t_j^i \log y_j^i + \lambda \|\theta\|^2$$



■节点 i 的预测情感分类概率分布向量 y^i : 每个节点利用基于其向量的 $softmax$ 分类器（如前所述）。

例如 y^i 为: $\{0.1, 0.1, 0.1, 0.1, 0.6\}$

■节点 i 的真实情感类分布向量 t^i : 采取0-1 编码的模式。即如果有5个类, 则其长度为5, 只有其中一个元素取值为1, 其它为零。

例如 t^i 为: $\{0, 0, 0, 0, 1\}$

- $E(\theta)$ 是关于参数 $\theta = (V, W, W_s, L)$ 的函数
- 调整参数, 以使交叉熵损失最小: $\theta_j := \theta_j - \alpha \cdot \frac{\partial E}{\partial \theta_j}$ 。

Thanks!

