基于神经网络的关系词非充盈态复句层次的自动识别

杨进才 杨璐璐 汪燕燕 沈显君
(华中师范大学计算机学院 武汉 430079)

摘 要 复句层次关系划分是复句句法结构分析以及语义甄别的基础,但关系词非充盈态复句由于关系标记的省略给层次划分带来了困难。文中利用依存关系句法树和 word2vec 词向量模型的方法来提取复句中分句的句法特征和语义特征,并利用神经网络进行训练,获得三句式关系词的非充盈态复句层次划分模型,对测试集中的复句进行层次划分测试,其准确率为 74%。

关键词 关系词非充盈态,复句层次划分,依存句法,word2vec,神经网络

中图法分类号 TP751 **文献标识码** A

Hierarchy Division of Compound Sentence with Non-saturated Relation Word via Neural Network

YANG Jin-cai YANG Lu-lu WANG Yan-yan SHEN Xian-jun
(School of Computer, Central China Normal University, Wuhan 430079, China)

Abstract Hierarchical division of a compound sentence is the basis of syntactic structure analysis and semantic discrimination. However, the ellipsis of relational markers bring difficulties to the hierarchical division of a compound sentence. This paper combined dependency syntactic trees and word2vec word vector model to extract the syntactic structure and semantic features of compound sentences, then used the neural network to train a hierarchy division model for compound sentences with non-saturated relation word, and the hierarchical division test was carried out on the complex sentences in the test set. The test accuracy of test set is 74%.

Keywords Compound sentence with non-saturated relation word, Hierarchical division of compound sentence, Dependency grammar, word2vec, Neural network

复句是汉语语法的重要实体,它包含两个或两个以上的分句[1],其中分句之间的层次结构和逻辑语义相对比较复杂,要理解复句的意思,必须先弄清复句内分句的语义及分句间的层次结构。汉语复句层次归属的正确划分对自动问答与信息抽取有着重要的意义,也有利于推动汉语复句的机器翻译,甚至是篇章理解的发展。

多重复句是包含不止一个层次结构的复句[2],其层次划分可以看作是两个以上分句之间相互选择,匹配成不同层级复句模块的过程。

关系标记(关系词)是复句中用来联结分句标明关系的词语,在多重复句中,它作为复句内部关系的标志,在复句中有着特殊的地位和作用。关系词的正确提取、标记和搭配是复句层次划分的重要依据[3-4]。关系词非充盈态复句关系标记的省略使得其层次划分一直都是复句层次划分的难点。因此,本文以三句式关系词非充盈态下的二重复句为研究对象。

例 1 ①如果片面追求升重点中学的升学率的问题不解决,②即使小学制延长到 6 年,③也不能解决负担过重的问题。(《人民日报》1984 年)

例 2 ①只要我们为人民的利益坚持好的,②为人民的利益改正错的,③我们这个队伍就一定会兴旺起来。(毛泽东《为人民服务》)

例 1 中,"即使……,也……"标明让步关系,"如果……[即使……,也……]"标明假设关系。两个句式嵌套起来,形成一个二重复句:如果……[即使……也……]"。例 1 的层次结构可以划分为:[①,[②,③]]。

例 2 中,分句②关系标记省略,所以仅依靠关系标记并不能识别出分句①和分句②之间的层次关系,无法划分例 2 的层次归属。但是,分句①和分句②之间存在"为人民的利益"这个共同的元素。此外,本体语义学对文本的理解就是对词汇语境捕获、结构化、形式化和认知计算的过程[5]。因此,需要从复句的分句中获取更多的信息来判别多重复句的层次归属。

1 相关工作

目前,对汉语复句层次划分问题的研究主要集中在有标复句,采用基于规则的方法。

鲁松等[6]提出层次关系树的概念,并以此为基础构造上下文无关文法表示多重复句,提出了一种基于预测机制,自底向上、部分数据驱动的确定性移进-规约算法来处理多重复句的层次关系。虽然该方法取得了很好的实验效果,但采用了

本文受国家社科基金项目(19BYY092)资助。

杨进才(1967—),男,博士,教授,博士生导师,主要研究方向为现代信息系统、中文信息处理,E-mail:jcyang@mail.ccnu.edu.cn(通信作者)。

基于构造规则的方法,模型的泛化能力不强。李幸等[7]在分析汉语标点符号用法和句法功能的基础上,提出一种新的面向汉语长句的层次化句法分析方法来判断多重复句的层次关系。该方法主要引入了标点符号来处理汉语长句的层次划分问题。

吴锋文[8-9]从主谓句法成分角度对分句关联进行深层知识挖掘,提取出直接聚层关联的分句间存在的 10 组典型特征,构建了一种基于分句语义关联度判定的复句分析法。胡金柱等[10]采用分句语义关联理论来解决无标复句以及关系词充盈态下某些特殊因果类、并列类复句的层次归属问题。从句法、语义两个角度,总结出分句语义关联的 3 大类,14 个小类的特征,提出了分句间语义关联度的计算方法,根据关联度来确定分句的层次归属。以上两种方法对无标复句的处理都有一定的意义,但主要从语言学角度出发去构建规则,但实验的数据集较小,不具有适用性。

李源等[11]提出了一种语义与规则相结合的复句层次分析模型,挖掘了影响分句关联的形式化语义知识,构建了小句关联体识别算法,并将其应用于复句层次判定规则之中,以辅助分析其层次关系。该方法的研究对象是有标复句。

本文的研究对象为三句式关系标记非充盈态下的二重复句。通过对分句进行依存语法分析以构建依存句法分析树来提取句法结构特征;对分句进行语义角色分析以计算相邻分句的语义特征并提取分句的主语特征。由于在特征的提取中并未涉及关系词,所以,本文方法适用于无标复句。

2 分类特征提取

2.1 分句句法特征的提取

依存语法是通过分析语言单位内成分之间的依存关系来揭示句法结构,该语法直接描述词语之间的关系。而多重复句的层次划分很大程度上受到语法成分的影响。因此,本文构建的依存句法分析树提取了句子中的“主谓宾”“定状补”等语法成分信息以及各成分之间的关系信息。复句中分句的句法特征提取过程如算法 1 所示。

算法 1 句法特征提取

输入:三句式关系标记非充盈态下的二重复句 S

输出:二重复句 S 分句间的依存句法相似度

1. 利用 Pyltp 对复句 S 中各分句进行依存语法分析,根据分析结果提取每个词的词性和该词与父节点的关系,根据词性和该词与父节点的关系构建依存句法分析树 T_i,$1\leqslant i\leqslant 3$。
2. 深度优先搜索算法遍历依存句法树 T_i 的节点,抽取分句 i 的路径,路径集合 $set_{T_i}=\{t_{i1},t_{i2},t_{i3},\cdots,t_{in}\}$,其中 n 代表分句 i 依存句法树的子路径个数,$1\leqslant i\leqslant 3$。
3. 遍历路径集合 set_{Ti} 中每一条路径 t_j。获取分句①和分句②的相同路径,添加到集合 $Tsame_{12}$ 中,$Tsame_{12}=set_{T1}\cap set_{T2}$;获取分句②和分句③的相同路径,添加到集合 $Tsame_{23}$,$Tsame_{23}=set_{T2}\cap set_{T3}$,其中 $1\leqslant j\leqslant n$。
4. 计算路径集合 $Tsame_{12}$ 中的路径个数,将其作为分句①与分句②的句法结构相似度 S_{sim12};计算路径集合 $Tsame_{23}$ 中的路径个数,将其作为分句②与分句③的句法结构相似度 S_{sim23},输出分句间的句法结构相似度。

2.2 分句语义特征的提取

语义角色标注是一种语义分析技术,即标注句子中的某些短语为给定谓词的语义角色,如施事、受事、时间和地点等。通过语义角色标注对复句内容进行语义处理,挑选重要的特征词语,脱离特征选择的繁复工作。如果两个特征词语的上下文很相似,那么在词向量空间中的位置也很接近[12]。例 3 中虽然没有明显的句法结构信息,但分句②中的“黄昏”和分句③中的“残秋”“黄昏”“平和”或“宁静”为上下文关系,在语义上很相似,所以利用词向量能更好地反映出词语的语义关联。

例 3 ①看起来谢王孙虽然并不太老,②可是他的生命却已到了黄昏,③就正如这残秋的黄昏般平和宁静。(古龙《三少爷的剑》)

复句中分句的语义特征提取过程如算法 2 所示。

算法 2 语义特征提取

输入:三句式关系标记非充盈态下的二重复句 S

输出:二重复句 S 分句间的语义相似度

1. 利用 Pyltp 对复句 S 中分句进行语义角色标注,提取分句的实施者、谓语、动作的影响者等特征词语。
2. 根据上一步提取的分句的特征词语,从训练的词向量模型(word2vec[13]在维基百科语料库上训练的 Skip-Gram 模型)[14-15]中提取词向量 x。
3. 计算句子的相似度,其计算式为:

$$Cosdis=\frac{(\sum_{j=1}^{k_1}\sum_{i=1}^{n}x_{ji})\times(\sum_{j=1}^{k_2}\sum_{i=1}^{n}y_{ji})}{\sqrt{\sum_{j=1}^{k_1}\sum_{i=1}^{n}(x_{ji})^2}\times\sqrt{\sum_{j=1}^{k_2}\sum_{i=1}^{n}(y_{ji})^2}} \tag{1}$$

其中,n 代表词向量的维度,k 代表特征词的个数。

4. 输出分句①与分句②的语义相似度 T_{ssim12},分句②与分句③的语义相似度 T_{ssim23}。

2.3 分句主语特征的提取

分句间的主语特征是否一致是判断复句层次归属的重要依据。利用语义角色标注功能对复句进行语义分析,使用主语特征提取器从语义分析结果中提取分句的实施者 A0,将分句的实施者 A0 作为分句的主语。例如,从图 1 中可以发现,分句①的实施者 A0 为“他”,分句②的实施者 A0 也为“他”,分句③的实施者 A0 为“你”,表明分句①和分句②有同一个主语。主语一致特征值设为 1,主语不一致特征值设为 0。

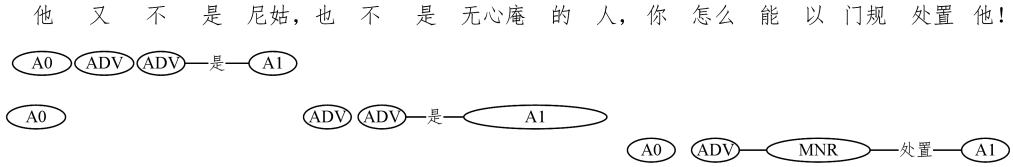

图 1 语义角色分析图

2.4　特征提取算法举例

例 4　①他又不是尼姑，②也不是无心庵的人，③你怎么能以门规处置他。

1. 依存句法特征的提取

(1)使用 pyltp 对例 4 的分句①进行分词、词性分析、依存句法分析。

分词结果：

['他','又','不','是','尼姑']

词性分析结果：

['r','d','d','v','n']

依存句法分析结果：

['4:SBV','4:ADV','4:ADV','0:HED','4:VOB']

索引 0 表示 root 节点，'是'的索引号为 4，'他'的父节点是'是'，'他'与父节点'是'的关系是 SBV；同样，其他词的父节点以及与父节点的关系以此类推。

根据词性分析和依存句法分析，以该词的词性和该词与父节点的关系为节点，对分句①构建一棵句法分析树(节点中的汉语词语只是为了便于理解，实际构建中不存在)，如图 2 所示。

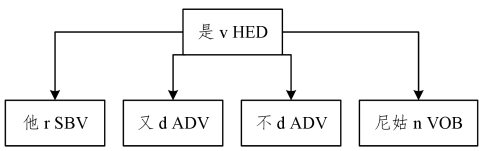

图 2　分句①的句法分析树

(2)同理，对例 4 中的分句②、分句③进行分词、词性分析、依存句法分析，并构建句法分析树，如图 3、图 4 所示。

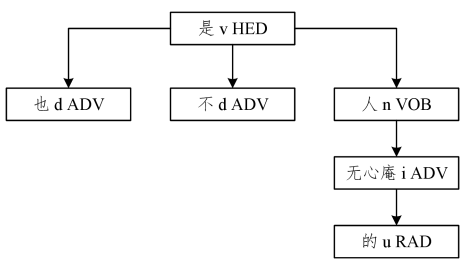

图 3　分句②的句法分析树

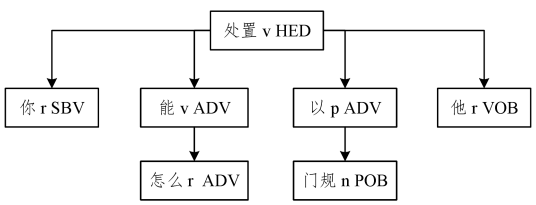

图 4　分句③的句法分析树

(3)提取分句①，②，③的路径集合。分句①的路径集合(包含 9 条路)：

1)['v','HED']

2)['r','SBV']

3)['d','ADV']

4)['d','ADV']

5)['n','VOB']

6)['v','HED'],['r','SBV']

7)['v','HED'],['d','ADV']

8)['v','HED'],['d','ADV']

9)['v','HED'],['n','VOB']

(4)分句①和分句②的相同路径：

1)['v','HED']

2)['d','ADV']

3)['n','VOB']

4)[['v','HED'],['d','ADV']]

5)[['v','HED'],['n','VOB']]

分句②和分句③的相同路径：

1)['v','HED']

(5)分句①和分句②的结构相似度：5

分句②和分句③的结构相似度：1

2. 分句语义特征的提取

(1)以例 4 中的分句③为例进行语义角色分析。

语义角色分析结果为：

5 A0:(0,0)ADV:(1,1)MNR:(3,4)A1:(6,6)

5 为谓词索引，即'处置'；A0 表示动作的实施者；A1 表示动作的影响。

(2)根据语义角色分析，提取分句的特征词语：

分句①的特征词语为：['他','是','尼']

分句②的特征词语为：['是','无心庵','的','人']

分句③的特征词语主为：['你','处置','他']

(3)从词向量模型中提取各分句特征词语的词向量 x。

(4)利用式(1)计算复句各分句间的语义相似度。

分句①与分句②的语义相似度：0.707426

分句②与分句③的语义相似度：0.563742

3. 分句主语特征的提取

分句①与分句②的主语特征为：1；

分句②与分句③的主语特征为：0。

4. 将句法特征、语义特征和主语特征输入到神经网络，来划分句子的层次结构。

3　神经网络模型分析

3.1　特征数据的分析

通过对提取的特征数据进行分析发现，很多特征数据对复句的层次划分贡献不明确。以表 1 的第一条数据为例，分句①与分句②的句法结构相似度为 3，语义相似度为 0.8064；分句②和分句③的句法结构相似度为 4，语义相似度为 0.7592，这些特征数据之间的差别不明显，且分句间的句法特征与语义特征不具有一致性。但通过分析发现，复句的层次与语义和主语联系更为密切，且主语特征为 0 的较多。因此，本文设计了一种带有数据加强机制的三层神经网络。

表 1　特征数据集

3	4	0.8064	0.7592	0	0
6	3	0.3059	0.6543	0	0
0	3	0.7739	0.6324	0	0
0	1	0.6230	0.5641	0	0

3.2　神经网络模型的构建

带有数据加强机制的三层神经网络模型如图 5 所示。在

网络的输入层前加一层强化数据层，通过对数据的强化使得特征数据层次划分更准确，同时也使得复句层次划分模型的泛化性能更好。数据强化层的原理是：主语特征 i=句法结构特征 $i*\alpha$+语义结构特征 $i*\beta$+主语特征 i，其中 $0<\alpha<\beta<1$，$i=1,2$。

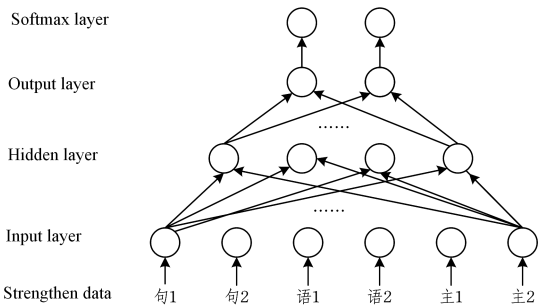

图 5 神网络模

神经网络模型算法的步骤如算法 3 所示。

算法 3 神经网络模型算法

输入：输入特征数据集 $D=\{((x_k,y_k))\}_{k=1}^{m}$ 学习率 η

过程：

1. 在(0,1)范围内随机初始化网络中所有连接的权重和阈值。
2. 将 $X_i=[x_1,x_2,x_3,x_4,x_5,x_6]$ 输入神经网络的数据强化层，主语特征发生变化，得到：

$$x_5=x_1*\alpha+x_3*\beta+x_5$$

$$x_6=x_2*\alpha+x_4*\beta+x_6$$

3. 将从强化数据层重新得到的特征数据 $X_i=[x_1,x_2,x_3,x_4,x_5,x_6]$ 输入神经网络的输入层。
4. 从输入层到隐含层进行一次非线性映射：

$$\varphi_j=f(X_i)=f(\sum_{i,j=1}^{i=6,j=4}w_{ij}x_i+b_j)$$

其中，隐含层使用的激励函数 f 是 ReLu，与 Sigmoid 和 Tanh 相比，Relu 能够快速收敛，可以更加简单的实现，而且有效缓解了梯度消失的问题。

5. 隐含层到输出层做一次线性变化，通过 Softmax 分类器得到不同类别的概率：

$$Y_l=\frac{\exp(\sum_{l=1}^{2}\sum_{j=1}^{4}v_{jl}\varphi_j+\theta_l)}{\sum_k\exp(Y_k)}$$

6. 实际输出值与期望输出值存在误差时，对 softmax 的输出向量和样本的实际标签做一个交叉熵损失函数，损失函数为：

$$loss=-\sum_l Y_k\log(Y_l)$$

Y_k 指实际标签中第 k 的值；Y_l 指 Softmax 的输出向量中第 l 个元素的值。

7. 本文使用了 Adam[16] 优化算法对模型进行训练，Adam 优化过程及权重更新规则为：

$$g\leftarrow\nabla\theta f_t(\theta_t)$$

$$v_t\leftarrow\beta_1\cdot w_{t-1}+(1-\beta_1)\cdot g$$

$$w_t\leftarrow\beta_2\cdot v_{t-1}+(1-\beta_2)\cdot g^2$$

$$\hat{v}_t\leftarrow v_t/(1-\beta_1^t)$$

$$\hat{w}_t\leftarrow w_t/(1-\beta_2^t)$$

$$\theta_t\leftarrow\theta_{t-1}-a\cdot\hat{v}_t/(\sqrt{\hat{w}_t}+\varepsilon)$$

其中，β_1 为一阶动量衰减系数，α 为步进值，β_2 为二阶动量衰减系数，ε 为数值稳定量，t 为 Timestep。

重复训练，直到误差小于设定的范围或达到指定的训练次数，得到神经网络的权值，输出神经网络模型。

4 实验分析与说明

本文从 CCCS 语料库(CCCS 语料库是由华中师范大学开发，复句语料均选自《长江日报》和《人民日报》)中挑选了 25000 个复句，从小说中选择了 5000 个复句。其中，测试集共10000 个，训练集 20000 个，语料分类统计情况如表 2 所列。

表 2 语料类别情况统计

层次关系	训练集	测试集
[①,[②,③]]	9501	4511
[[①,②],③]	10499	5489

为了验证带有数据加强机制神经网络模型的有效性，测试了 3 种模型。支持向量机(SVM)和随机森林被广泛使用且分类效果较好。支持向量机是一个能够将不同类样本在样本空间分隔的超平面；随机森林是一个组合分类器，其利用随机生成的决策树对样本进行训练并预测。3 个模型在相同的训练集和测试集中从曲线、准确率、Auc、f1 值、召回率几个方面进行对比，如表 3 与图 6 所示，结果表明相比于 SVM、随机森林传统方法，本文方法在分类准确率方面得到了有效提高；在精确率与召回率指标上也取得了更佳的效果。

表 3 测试结果

算法模型	Accuracy	F1-score	Recall
神经网络	0.741	0.709	0.747
SVM	0.729	0.646	0.586
随机森林	0.707	0.666	0.692

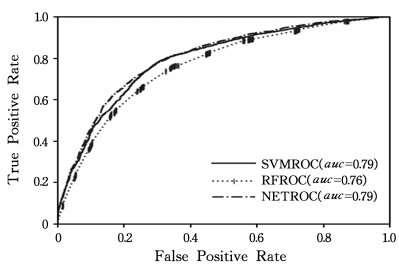

图 6 3 种模型的 roc 曲线

其中分类错误的原因有以下几点：

(1)汉语言本身非常复杂，有些复句仅靠句法和语义也很难划分，需要引入语境来分析。例如针对三句式无标并列句的层次划分。

例 5 ①你不觉得这太突然了吗，②你不觉得这根本不可能吗，③你不觉得这事太离谱了吗？

例 5 的分句①和分句②的句法相似度为 4，语义相似度为 0.915677；分句②和分句③的句法相似度为 4，语义相似度为 0.938668，主语一致。模型划分的层次结构为[①,[②,③]]，但因为例句 5 是并列复句，所以一般情况下划分的层次结构是[[①,②],③]

(2)特征提取时，使用的工具分析复句时存在一定的错误，有待进一步改善。例如：在进行语义角色分析时，有时会忽略最后一个分句，提取的特征不完整。特征数据的不准确或不完整会影响模型的准确率，从而造成层次划分

错误。例如，对例句 6 进行语义角色分析时，分句③会被忽略，如图 7 所示。

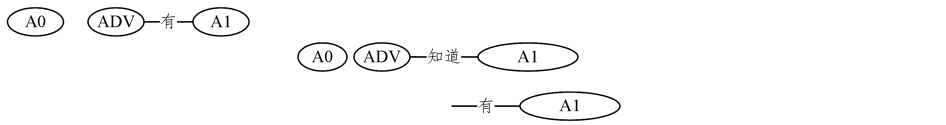

图 7　语义角色分析

例 6　①每个人都有潜意识，②当你不知道有冰淇淋的时候，③你会心甘情愿的喝水。

结束语　正确划分关系词非充盈态下复句的层次结构，须充分利用复句的结构信息和语义信息。依存句法树包含了句子的大量结构信息，而 Word embeddings 包含了大量语义信息。本文充分利用句子的结构信息与语义信息，提出了一种具有数据强化机制的三层神经网络模型，对三句式关系词非充盈态下二重复句的进行层次划分，获得了 74% 的准确率，这表明了该方法的有效性。

复句层次结构的判定需要综合考虑其句法、语义、关系词、语境等特征。而 RNN 和 CNN 在文档情感分类、文本分类都取得很好的效果[17-18]，下一步的研究工作是如何利用 RNN，CNN，LSTM 等深度学习的方法提取复句的特征，更加有效地划分关系词非充盈态复句的层次结构。

参 考 文 献

[1] 邢福义. 汉语复句研究[M]. 北京：商务印刷馆，2003

[2] 李晋霞，刘云. 面向计算机的二重复句层次划分研究[C]// 第 7 届计算语言学联合学术会议论文. 2003

[3] 胡金柱，陈江曼，杨进才，等. 基于规则的连用关系标记的自动标识研究[J]. 计算机科学，2012，39(7)：196-200.

[4] 杜超华，胡金柱，沈威，等. 基于复句语料库分词系统研究 [J]. 计算机数字与工程，2007，35(5)：43-44.

[5] 王飞，易绵竹，谭新. 基于本体语义网络的语言理解模型[J]. 计算机科学，2018，45(S1)：101-105.

[6] 鲁松，白硕，李素建，等. 汉语多重关系复句的关系层次分析[J]. 软件学报，2001，12(7)：987-995.

[7] 李幸，宗成庆. 引入标点处理的层次化汉语长句句法分析方法[J]. 中文信息学报，2006，20(4)：8-15.

[8] 吴锋文. 基于主谓语知识挖掘的分句语义关联研究[J]. 语言文字应用，2011，20(4)：132-142.

[9] 吴锋文. 面向信息处理的“一标三句式”复句层次关系判定[J]. 北方论丛，2012，54(1)：64-68.

[10] 胡金柱，舒江波，罗进军. 汉语复句中分句的语义关联特征[J]. 语言文字应用，2010，19(4)：121-130.

[11] 李源，等. 基于语义与规则的有标复句层次体系研究[J]. 计算机工程与科学，2017，39(12)：2306-2313.

[12] 黄仁，张卫. 基于 word2vec 的互联网商品评论情感倾向研究[J]. 计算机科学，2016，43(S1)：387-389.

[13] MIKOLOV T，SUTSKEVER I，CHEN K，et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. 2013：3111-3119.

[14] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[EB/OL]. http://arxiv. org/pdf/1301. 3781v3. pdf.

[15] YU M，DREDZE M. Improving lexical embeddings with semantic knowledge[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore，Stroudsburg，USA：ACL，2014：545-550.

[16] KINGMA D P，JIMMY B A. Adam：A Method for Stochastic Optimization[C]// 3rd International Conference for Learning Representations. San Diego，2015.

[17] TANG D Y，QIN B，LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015：1422-1432.

[18] LAI S，XU L，LIU K，et al. Recurrent convolutional neural networks for text classification[C]// Proceedings of the National Conference on Artifical Intelligence. 2015：2267-2273.