基于 TextRank 和字符级卷积神经网络的小学作文素材自动分类模型研究

朱晓亮 石昀东

430079

摘 要 随着教育技术与信息技术的融合,实现面向小学生的语文写作自动辅助成为可能。快速自动地进行范文素材的分类入库是实现写作自动辅助的关键。作文素材语义信息丰富、种类较多,若采用现有方法进行自动分类入库操作往往难以取得好的效果。因此,在分析小学作文的类别特征并构建了一个数据集的基础上,提出基于 TextRank 和字符级卷积神经网络的小学作文自动分类模型。运用基于 TextRank 的关键句提取模型为范文素材,去除部分冗余的语义信息。应用 word embedding 对数据集进行文本表示,并将其作为卷积神经网络的输入。通过不断地迭代训练和测试,最终实现了该模型。实验表明了该方法对于作文分类任务能显著地提高分类的性能。

关键词 TextRank 卷积神经网络 作文素材库 文档分类

中图分类号 TP183 **文献标识码** A DOI 10.3969/j.issn.1000-386x.2019.01.040

AUTOMATIC CLASSIFICATION MODEL OF COMPOSITION MATERIAL IN PRIMARY SCHOOL BASED ON TEXTRANK AND CHAR-LEVEL CNN

Zhu Xiaoliang Shi Yundong

National Engineering Research Center for E-learning Central China Normal University Wuhan 430079 Hubei China

Abstract With the integration of education technology and information technology it is possible to realize the automatic guidance of composition for primary school students. Fast and automatic classification and storage of model materials is the key to achieve automatic guidance of writing. Composition materials are rich in semantic information and various. It is often difficult to achieve good results via normal methods for automatic classification and storage. Therefore on the basis of analyzing the category features of compositions in primary school and constructing a data set an automatic classification model for compositions in primary school was proposed based on TextRank and character-level CNN. The key sentence extraction model based on TextRank was adopted to remove some redundant semantic information for the essay materials. The word embedding was applied to express the text of data set and took it as the input of convolutional neural network. The model was realized through continuous iterative training and testing. Experimental results show that this model can obviously improve the performance of composition classification.

Keywords TextRank CNN Composition library Text classification

1 - 2

0 引 言

2018 - 07 - 23 2018YFB1004504 18YJAZH152
CCNU18TS005 朱晓亮 石昀东

F1-score　84.41%

TextRank

TextRank

word embedding　word embedding

TextRank

1

2

3　TextRank

4

TextRank

5　TextRank

1　相关研究

[3]

– TF-IDF

LDA

VSM [4]

word embedding　Word2Vec　word embedding

Hinton [5]　Word2Vec　Tomas Mikolov [6]

[7]

Logistic

SVM

[8]

9

word embedding

10

[11]　LDA　word embedding [12]

Word2Vec　TF-IDF

13　doc2vec

Word2Vec

[14]　Word2Vec

[15]　word embedding

[16]

2　小学各类作文特征分析

[17]

2.1　写景类作文

“　”

[18]

2.2 写物类作文

19

2.3 写人类作文

20

2.4 记事类作文

21

2.5 应用文

17

22

1

表 1 小学各类作文的特征

3 理论与模型

3.1 TextRank 算法

TextRank 23 2004

PageRank

LDA HMM TextRank

$G = V\ E$

TextRank V E

E $V \times V$ $V_i\ V_j$

w_{ij} $V_i\ In\ V_i$

$Out\ V_i$ V_i V_i

$$WS\ V_i = 1 - d + d \times \sum_{V_i \in In\ V_i} \frac{w_{ij}}{\sum_{V_k \in Out\ V_i} w_{jk}} WS\ V_j \quad 1$$

d 0 1

0.85

1 TextRank

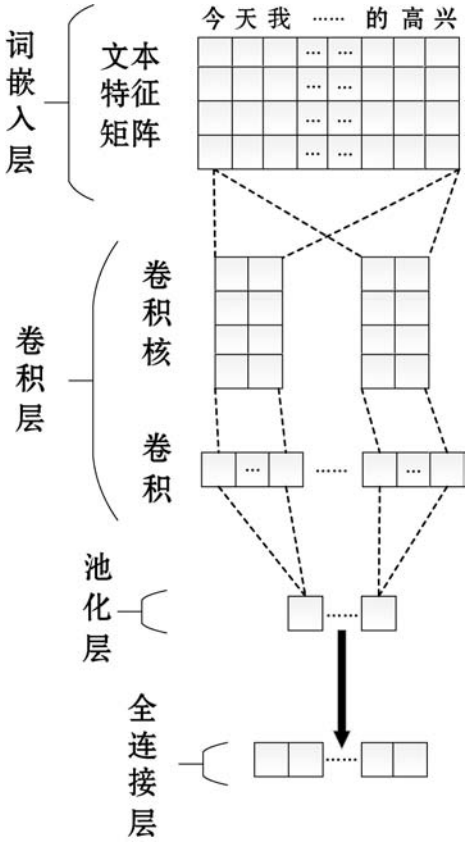
0.000 1

TextRank

3.2　卷积神经网络分类器

Lecun [24] 1989

1



1

embedding

word embedding

word embedding

$\boldsymbol{T} \in R^{d \times v}$

d　　v　word embedding

SGD　word embedding　word embedding [10]

$C \in R^{h \times v}$　h　$\boldsymbol{T}$

v　h　v

r_i　$i = 1$　2　…　$d - h + 1$

$$r_i = \boldsymbol{C} \times \boldsymbol{T}_{i\ i+h-1} \quad (2)$$

r_i　b

$f(x) = \max(0, x)$　ReLu　s_i　$i = 1$　2　…　$d - h + 1$

$$s_i = \max(0, r_i + b) \quad (3)$$

$d - h + 1$

$S \in R^{d-h+1 \times 1}$

m　S_j　$j = 1$　2　…　m

$$S_j = [s_1\ s_2\ \ldots\ s_{d-h+1}] \quad (4)$$

$\max(S_j)$

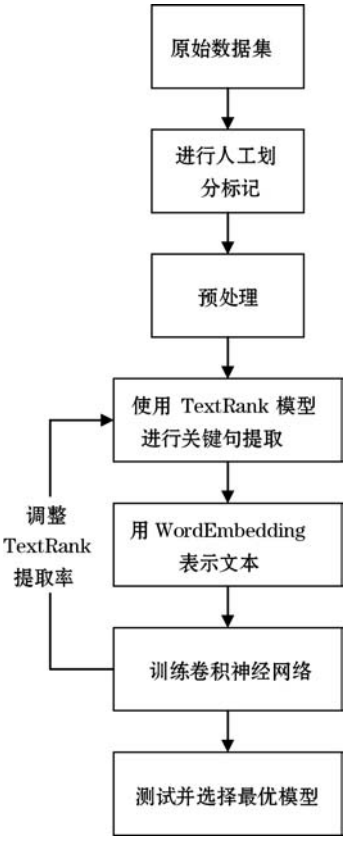
$P \in R^{m \times 1}$

P　softmax　$\boldsymbol{P}$

3.3　基于 TextRank 和字符级卷积神经网络的作文自动分类模型

TextRank

2



2

TextRank

word embedding

TextRank

4 实 验

4.1 实验环境

2

表 2 实验环境

	Ubuntu16.04.3
Cpu	IntelXeonE7 - 4820V31.90 GHz
	32 GB
	Python3.6.3
	Tensorflow1.5.0

4.2 数据收集和预处理

16 415

1 3 000

15 000

15%

2 250 10

9 1

10

10

4.3 评价指标

F1 F1-score

F1 F1-score = 2 × precision × recall / precision + recall

precision = /

recall = /

100

120

80

precision = 80/120 = 66.67%

recall = 80/100 = 80.00%

F1 F1-score = 2 × 66.67% × 80.00%/66.67% + 80.00% = 72.71%

F1

F1-score

4.4 实验设计

TextRank

5

3

表 3 超参数设置

word embedding	128
	1 000
	256
	1e - 3
	5
	128
Dropout	0.5

实验 1 TextRank

TextRank

实验 2

TextRank

实验 3

实验 4

实验 5 TextRank

1

word embedding word embed-

ding　1　word embedding

4.5　实验结果

4

表 4　实验结果

		F1-score
Char-level CNN + TextRank	0.6	87.82%
Char-level CNN + TextRank	0.7	88.47%
Char-level CNN + TextRank	0.8	89.24%
Char-level CNN + TextRank	0.9	88.45%
Char-level CNN	–	88.12%
Word-level CNN	–	84.41%
Char-level CNN + Title	–	79.67%
Knn + TextRank	0.8	25.22%
Logistic regression + TextRank	0.8	77.10%
Random forest + TextRank	0.8	28.55%
Decision tree + TextRank	0.8	28.92%
Linear SVM + TextRank	0.8	76.53%

4.6　结果分析

4.6.1　不同的 TextRank 提取率

4　5　TextRank

0.6　1.0

1.0

0.8

TextRank

4

0.8

4.6.2　以题目作为数据集

“　”

“　”

TextRank

F1-score　8.57%

4.6.3　词级卷积神经网络

4

F1-score

4.83%

4.6.4　传统分类模型

4

Logistic regression　F1-score

77.10%　F1-score

12.14%

4.6.5　其他影响因素

“

”

5　结　语

TextRank

F1-score　89.24%

“　12.14%

”

参 考 文 献

1 . J . 2017 11 101.

2 . J . 2012 28 5 495 - 501.

3 . J . 2004 18 1 26 - 32.

4 . D . 2016.

5 Hinton G E. Learning distributed representations of concepts C // Proceedings of the Eighth Annual Conference of the Cognitive Science Society. Amherst USA Eighth Annual Conference of the Cognitive Science Society 1986 1 - 12.

6 Mikolov T Chen K Corrado G et al. Efficient Estimation of Word Representations in Vector Space EB . eprint arXiv 1301.3781 2013.

7 Lilleberg J Zhu Y Zhang Y. Support vector machines and Word2vec for text classification with semantic features C // 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing. IEEE 2015 136 - 140.

8 . M . 2016.

9 Kim Y. Convolutional neural networks for sentence classification C //Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP 2014 . Doha Qatar EMNLP 2014 2014 1746 - 1751.

10 Zhang X Zhao J Lecun Y. Character-level Convolutional Networks for Text Classification C //Proceedings ofthe 2015 Neural Information Processing Systems Conference NIPS 2015 . Montreal Canada NIPS 2015 2015.

11 . LDA Word Embedding D . 2016.

12 . Word2Vec J . 2017 1 57 - 62.

13 LeQ Mikolov T. Distributed representations of sentences and documents C //Proceedings of the 31th International Conference on Machine Learning ICML 2014 Beijing China 2014 1188 - 1196.

14 . J . 2018 2 644 - 650.

15 . word embedding CNN J . 2016 33 10 2902 - 2905.

16 . J . 2017 34 4 157 - 164 177.

17 . J . 2011 34 33 - 33.

18 . — J . 2012 7 42 - 42.

19 . J . 2016 22 102.

20 . J . 2015 3 99 - 100.

21 . J . 2014 12 157.

22 . J . 2016 5 38 - 41.

23 Mihalcea R Tarau P. TextRank Bringing Order into Texts C //EMNLP 2004 404 - 411.

24 Lecun Y Bottou L Bengio Y et al. Gradient-based learning applied to document recognition J . Proceedings of the IEEE 1998 86 11 2278 - 2324.

(上接第 151 页)

参 考 文 献

1 · . Oracle Database 12c DBA M . . 2016.

2 Whalen E Czuprynski J. Oracle M . . 2017.

3 Freeman R G Hart M. Oracle Database 12c Oracle RMAN M . . 4 . 2017.

4 . PowerBuilder M . 4 . 2013.

5 . PowerBuilder M . 2011.

6 . PowerBuilder M . 2010.

7 . PowerBuilder M . 2010.

8 . PowerBuilder 10.5 M . 2009.

9 . PowerBuilder M . 2012.