中文作文句间逻辑合理性智能判别方法研究

刘 杰 孙 娜 袁克柔 余笑岩 骆力明
(首都师范大学信息工程学院 北京 100048)

摘 要 在作文评测中,句间逻辑合理性是评价语言运用能力的一项重要指标。从文本分类的角度,对作文段落句间逻辑合理性进行定性分析。依据逻辑合理的段落其句子的位置是相对固定的,将现有的基于传统、基于深度学习的文本分类算法应用在中小学人物类作文段落句间逻辑合理性的判别上,实验结果表明使用分类模型对段落句间逻辑合理性判别是有效的。在此基础上,进一步提出代表句子特征的关键词筛选方法,实验结果表明使用句首关键词、句尾关键词或两者结合作为句子特征的方法,比句子全部关键词更能代表句子信息,明显提高判别的准确率。

关键词 作文评测 BiRNN 句间逻辑合理性 无监督学习 文本分类

中图分类号 TP181 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.01.013

INTELLIGENT DISCRIMINANT METHOD OF LOGICAL RATIONALITY BETWEEN SENTENCES IN CHINESS COMPOSITION

Liu Jie Sun Na Yuan Kerou Yu Xiaoyan Luo Liming
(*College of Information and Engineering*, *Capital Normal University*, *Beijing 100048*, *China*)

Abstract In composition evaluation, logical rationality between sentences is an important indicator to evaluate the ability of language application. From the perspective of text classification, this paper made a qualitative analysis of the logical rationality between sentences of paragraphs. According to the fact that the position of the sentence was relatively fixed in a logical and reasonable paragraph, we applied the existing text classification algorithms based on traditional and deep learning into the discrimination of logical rationality between sentences of paragraphs in character composition in primary and secondary schools. The experimental result shows that it is effective to use the classification model to distinguish logical rationality between sentences of paragraphs. On this basis, the keyword selection method representing the features of the sentence was further proposed. The experimental result shows that the method of using sentence heading keywords, sentence ending keywords or the combination of the two as the sentence features can represent the sentence information better than all the sentence keywords, which significantly improves the accuracy of discrimination.

Keywords Composition evaluation BiRNN Logical rationality between sentences Unsupervised learning Text classification

0 引 言

作文写作可以考察考生的逻辑思维与语言运用的能力和水平,是汉语考试中必考科目。国内中考、高考考生和中国汉语水平考试的考生数目十分巨大,如此大规模的考生数量,抽调大量人员参加阅卷任务,人工成本极高,耗费大量物力和财力。中文作文的自动评分研究逐渐兴起,对于提高评分效率,从根本上消除作文评价的不一致,控制评分误差具有十分重要的意义。由于中文语言逻辑的复杂程度大,现有的研究对作文评测大多从词汇使用、语法表达、作文长度、关联词使

收稿日期:2018-07-07。国家自然科学基金项目(61371194,61672361)。**刘杰**,教授,主研领域:自然语言处理,语义网。**孙娜**,硕士生。**袁克柔**,硕士生。**余笑岩**,本科生。**骆力明**,教授。

用、修辞手法的运用、文章主题一致性等角度进行评测,并未涉及作文内部逻辑合理性评测。但是在作文评测中,逻辑合理性同样是评价语言运用能力的一项重要指标。因此,本文从逻辑结构这个方面考虑,选取段落逻辑组织结构作为研究点,研究段落中句子组织的可读性,从而达到从逻辑的角度,考察学生语言运用水平,辅助作文逻辑合理性评测,为更加全面的进行作文自动评测提供了帮助。

1 相关研究

现有的中文文本逻辑的研究,主要涉及对长文本进行逻辑上的划分。长文本具有一定的文本结构,可以表达更多的语义,因此将长文本按照逻辑分块,能够更好地分析文本生成的过程以及文本内部语义的连贯性,挖掘文本内部的逻辑关系。

文献[1]将连贯的文本认为是一种文本生成流,专注于信息流的文本分割,提出依据文本主题进行分割。傅间莲等[2]提出建立段落向量空间模型,通过对段落相似度的度量实现对文本主题的划分,使文本的内容更加全面,篇章的结构更加严谨,对文本结构以及逻辑是否合理有一定的指导作用。

除了上述对长文本进行逻辑划分,分析文本内部语义的连贯性的方法外,还可以用连接并表明文本间关系的关联词来评判中文文本间逻辑。比较有代表性的概念层次网络(HNC),是以语义表达为基础的自然语言理解处理技术,是中科院声学所黄曾阳研究员创立的面向整个自然语言理解的理论框架[3-6]。HNC 理论认为语句从属的类别能够从宏观角度表示该语句,对句类划分为 57 个基本句类并用字母符号组合为每个句类生成表达式,反映 7 种类型的语句;而其内部的语义块之间具有特定的概念关联性,能够从微观层面表征该语句,对语义块内语句的逻辑顺序做了研究。另外,司贝贝等[7]通过汉语依存句法,挖掘语料中关系词及其搭配信息,自动识别并提取出文本中的关系词,建立关系词搭配语料库,为分析语句中关系词搭配规律提供数据上的支持。李艳翠等[8]构建的汉语篇章结构分析平台,依赖对语料进行包含子句、连接词、篇章关系和篇章结构树等方面的标注,自底向上进行汉语篇章结构分析,识别子句、连接词、篇章关系以及完成篇章结构树构建等任务。杨进才等[9]通过广泛分析汉语复句语料库,从关系词在文本中的环境位置和相关的组合搭配方面提取必要的特征,使用贝叶斯方法对上述特征进行训练和测试,挖掘文本中语义上的逻辑关系,具有可行性和有效性。

上述研究方法需要大量的标注数据,另外,考虑到语义上有所关联的复句,除了依靠关联词之外,更多的文本间的联系则是通过隐含语义来衔接的。本文以实际应用为驱动,拟完成对中小学作文段落整体逻辑合理性的判别,在语料上缺乏大量的人工进行标注分析,且处理对象为具有实际意义的段落整体,而非简单的语义逻辑块。因此,本文拟通过无监督的方式,减少人工成本,达到对段落逻辑合理性定性分析的目的,将对段落内容逻辑的划分转为对段落逻辑合理性判别的分类问题。

2 判别任务及模型确定

2.1 数据来源与判别任务

本文以实际应用为前提,从逻辑结构方面考虑,选取段落逻辑组织结构作为研究点,拟完成对中小学作文段落整体逻辑合理性的判别,考察学生语言运用水平,对文本逻辑结构方面进行判别,最终为作文自动评测提供帮助。

小学阶段主要涉及记叙类作文,范围包括了写景、写物、写事和写人四个方面。除此之外,还有一些想象类作文、情感类作文、应用类作文及话题作文等等,但相对较少。中学作文体裁还包含了除记叙文之外的说明文、读后感等形式。因此,本文在选取语料方面,最终选取了记叙类作文中的人物作文作为验证依据,从网络上各个作文网站获取中小学人物类作文训练语料 11 766 篇,测试语料 4 563 篇,总计 16 329 篇。

每篇作文由不同的段落构成,相对于对段落内部进行小语义块或复句的识别,从而评价语义块逻辑,本文则选取段落整体作为研究对象,从整体上对其进行合理性的判别。通过分析,采用了两种分类任务对文本逻辑合理性进行判别,其各自的依据及对应的判别任务如下:

第一,由于段落内部,每句话都有相应的位置,且不可轻易调换,因此,本分类任务拟通过判断段落内部每句话是否在相应的位置,从而判断整个段落的结构是否合理,构建段落内部句子-位置判别模型 S2PM(Sentence to Position Model)。

第二,逻辑合理的段落,其句子的顺序是一定的,若其内部的句子顺序改变,几乎一定会影响到文本整体的连贯性,依此构建段落整体逻辑合理性判别模型 LRDPM(Logical Rationality Discrimination of Paragraph

Model,简写为 DPM)。

2.2 模　型

中文不同于英文,单词之间并没有空格来间隔。任何中文的数据预处理阶段都包含分词阶段,分词的准确性和有效性直接影响着数据预处理的效果,而且对之后文本进一步处理有很大的影响。相比结巴分词、中科院自主研发的 ICTCLAS 中文分词系统,哈工大社会计算与信息检索研究中心历时十年研制的语言技术平台 LTP(Language Technology Platform)有更强的领域适用性[10-11],是目前分词的主流方法,因此本文采用的分词工具为 LTP。

2.2.1 算法的选取

S2PM 与 DPM 都依赖分类算法完成相对应任务。本文从传统机器学习算法以及深度学习算法中选取了朴素贝叶斯、支持向量机 SVM、卷积神经网络 CNN、双向递归神经网络 BiRNN(Bi-directional RNN)、fastText 五种分类算法,完成 S2PM 与 DPM 两个模型的任务。

首先,朴素贝叶斯在贝叶斯的基础上,加入条件独立的假设,通过观察数据本身,得到类别先验概率、给定类别下观察到不同数据的条件概率,直接且高效地计算待测数据属于某个类别的后验概率[12-13]。朴素贝叶斯模型发源于古典数学理论,对缺失数据不太敏感,算法也比较简单,常用于文本分类,有稳定的分类效率。SVM 能够最小化结构风险,具有较好的泛化推广能力,且其在模式识别领域中的文本识别、人脸识别等方面应用也取得不错的效果[14-16],因此纳入本文的基础分类算法之一。

其次,深度学习作为机器学习中一种基于对数据进行表征学习的方法,在越来越多的领域发挥重要的作用,因此本文加入了深度学习的一些主流经典算法。例如,递归神经网络 RNN(Recurrent Neural Network)能够更好地表达上下文信息,应用在序列标注、命名体识别、seq2seq 等方面[17],在文本分类任务中,还存在 BiRNN 捕获变长且双向的"N-gram"信息;CNN 是利用卷积神经网络对文本进行分类的算法,CNN 和 N-gram 模型相似:CNN 中的 filter 窗口可以看作是 N-gram 的方法,因而可以被运用到文本分类中[18-19],CNN 使用的卷积层和 pooling 层,使得训练过程中参数个数减少的同时,还能够抽取到文本的更高层的信息;fastText 相比深度学习,训练速度极快, Mikolov 把 fastText 应用在文本分类上,在训练的过程中,不仅可以得到文本内每个词对应的词向量,解决文本表征的问题,还能得到分类模型,且分类效果在一些数据集上接近甚至超过了 CNN、RNN[20-21]。

综上所述,本文最终选取了朴素贝叶斯、SVM、CNN、BiRNN、fastText 五种分类算法,完成 S2PM 与 DPM 两个模型的任务。

2.2.2 特征及输入输出表示

在构建 fastText 分类模型的过程中,词向量字典作为其附属也随之生成,因此本文将其生成的词向量,作为词语的分布式表示特征,参与到其他各个分类器的分类任务中。

任务一拟通过判断段落内部每句话是否在相应的位置,从而判断整个段落的结构是否合理,构建段落内部句子-位置判别模型 S2PM。其输入为段落内部的句子,由句子内词语的分布式表示特征加和取平均作为句子的向量化表示方式,而其在段落中的位置作为其类别标签。任务二构建了段落整体逻辑合理性判别模型 DPM,将整个段落作为输入,此时,段落则可表示为句子向量的拼接,而句子的向量化表示方式仍为词语的分布式表示特征加和取平均。若输入段落内的句集顺序正常,则对应标签为合理;若输入的句集顺序混乱,则对应标签为不合理。

2.2.3 模型设计

本文构建相关分类训练以及测试模型,如图 1 所示。

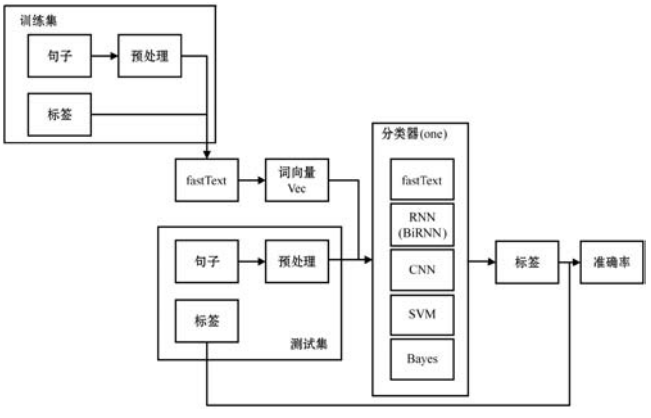

图 1　作文段落逻辑合理性判别模型

在以上两类任务中,本文均首先选取了中小学人物类作文训练语料 11 766 篇,生成关于人物类作文领域词语的词向量。其次,测试语料为 4 563 篇,采用 2.2.2 节所介绍的输入和输出表示方式,将不同的分类算法应用在同一测试数据集上。最后,分析分类器预测标签与文本实际标签的差距,从而评判模型效果。

3　实验结果与分析

本节介绍了实验所需的数据预处理过程,在处理不同的分类任务中,对比不同算法的表现效果,并对其原因进行了深入分析。

3.1 数据预处理

本文从网络上各个作文网站获取中小学人物类作文训练语料 11 766 篇,测试语料为 4 563 篇,总计 16 329 篇。

本文预先进行了段落内部句子长度的统计。图 2 给出了段落内部句子长度在 2 ~ 40 之间的段落数量,40 句以上的每类段落的总数均小于 500,因此未在图中给出。从图 2 可以发现,段落数量在 1 500 及以上的,其内部句子长度在 3 ~ 23 之间。由于段落内句子越多,其内部存在的关系越复杂,并且为了保持段落内部语义的完整性,不考虑将长句切分为短句,本文最终决定从段落内部句子长度在 3 ~ 23 之间的前 1/3 左右的段落,即段落内部句子数量在 2 ~ 7 之间的段落作为研究对象。

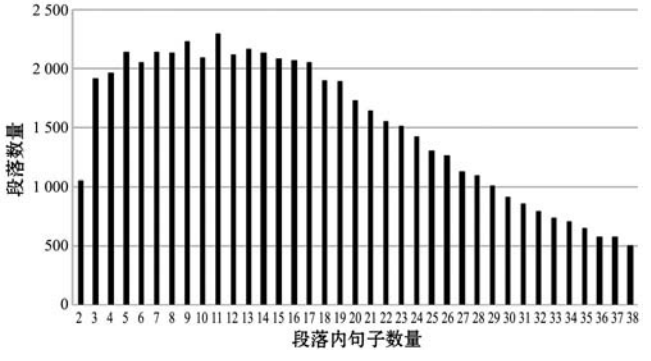

图 2 段落句子数量与段落数量分布图

句子中并非完全都是有意义的词语,本文在使用 LTP 对每种段落内词语的词性标注的同时,对各词性数量进行了统计,如表 1 所示。有意义的词性中,动词(v)、名词(n)、形容词(a)居多,因此,在处理语料时,仅抽取了上述三类词性的词语作为句子特征,而无意义的词,例如标点符号(wp)、停用词(u)等或者数量较少的词性作为无关特征将被过滤。

表 1 不同句子数量的段落内部词语词性统计

词性	句长					
	2	3	4	5	6	7
v	27 329	44 944	34 084	28 246	18 574	11 162
wp	23 051	36 453	26 475	21 589	14 273	8 123
n	21 542	34 120	23 429	19 144	12 351	7 195
u	14 986	22 598	16 612	13 753	9 237	5 596
r	14 226	21 762	16 566	13 523	9 032	5 586
d	11 522	17 556	13 777	11 440	7 775	4 575
a	7 617	11 341	7 967	6 360	4 152	2 484
m	5 234	8 546	5 266	4 274	2 818	1 648
p	4 369	6 777	5 188	4 213	2 612	1 604
q	3 708	6 087	3 829	3 002	1 887	1 115

3.2 S2PM 实验效果

本节按照图 1 的模型,对中小学人物作文段落内部的句子进行位置判断,拟通过判断句子是否在其段落内的原始位置来评判段落逻辑的合理性。分类结果对比如图 3 所示。

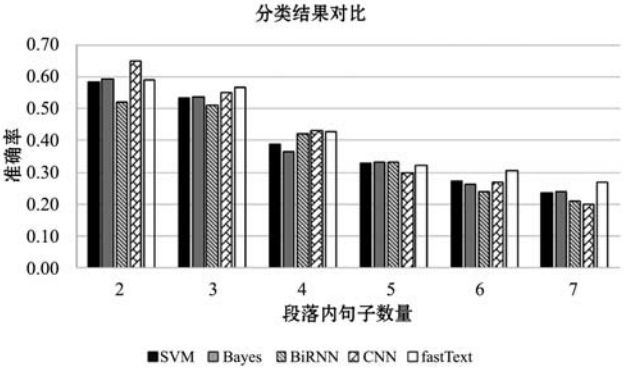

图 3 句子-位置模型判别结果

图 3 给出了五种方法构建句子-位置判别模型,综合考虑,在段落长度为 2 的情况下,CNN 对句子在段落中的位置预测准确率最高,接近 65%;其余情况下,fastText 对预测的结果是所有模型中表现最佳的。传统的 SVM 与朴素贝叶斯同深度学习模型的表现十分接近,甚至在不同的段落句长下分类准确率超过了 BiRNN 与 CNN。

3.3 S2PM 实验结果分析

从上述实验结果来看,本文构建的 S2PM 表现效果一般,最高准确率仅达 65%,并不理想,因此本节对此进行了深入分析。构建 S2PM 任务的提出是基于每个句子的特征,即关键字,由于句子在段落中位置不同而存在一定的区分度。卡方检验用途非常广,用来分析样本的实际观测值与理论推断值之间的偏离程度,用 χ^2 表示,实际观测值与理论推断值之间的偏离程度决定卡方值的大小,例如检测关键词与类别之间的相关性大小,卡方值越大,表明两者越不符合;卡方值越小,偏差越小,两者越趋于符合;若两个值完全相等,卡方值就为 0,表明两者完全符合。关键词与类别为一对一的映射关系。由于关键词数量过多,若采用对于 $R \times C$ 的列联表的卡方检验,仅在自由度 $v = (R-1)(C-1)$ 的计算上就超过了查 χ^2 临界值表所涉及的范围。因此,本节选用四格卡方表,具体实现过程如下:

首先,建立假设,原假设 H_0 为特定关键词 W 与特定类别,即句子所在段落的位置无关,条件假设 H_1 为特定关键词与特定类别相关;$\alpha = 0.05$,查 χ^2 临界值表,在自由度 $v = (R-1)(C-1) = 1$ 的情况下,认为 $\chi^2_{0.05}(1) > 26.3$ 则拒绝原假设,关键词与其句子所在段落位置是相关的,反之则支持原假设。根据统计,如

表 2 所示。

表 2　关键词-类别统计示例图

关键词	类别	其他类别	总数
关键词 W	$A(T_{1.1})$	$B(T_{1.2})$	$A+B$
其他关键词	$C(T_{2.1})$	$D(T_{2.2})$	$C+D$

表 2 中的 T 代表 TRC 理论数的缩写,表示第 R 行 C 列格子的理论数。计算公式如下:

$$TRC=\frac{nR\times nC}{n} \tag{1}$$

nR 为理论数同行的合计数,nC 为与理论数同列的合计数,n 为总例数(即 $A+B+C+D$)。根据表中统计,计算 χ^2 如下:

$$\chi^2=\sum\frac{(A-T)^2}{T}=\frac{(A-T_{1.1})^2}{T_{1.1}}+\frac{(B-T_{1.2})^2}{T_{1.2}}+\frac{(C-T_{2.1})^2}{T_{2.1}}+\frac{(D-T_{2.2})^2}{T_{2.2}} \tag{2}$$

表 3 选取了前 20 个由上述方法计算得到的关键词与句子位置之间的关系。

表 3　句子位置与相关关键词

句子位置	2	3	4	5	6	7
关键词	妈妈	老师	老师	老师	妈妈	妈妈
	老师	妈妈	妈妈	妈妈	眼睛	眼睛
	爸爸	爱	爸爸	会	爸爸	爸爸
	看	爸爸	爱	好	爱	爱
	去	看	眼睛	爸爸	看	看
	爱	要	看	眼睛	要	老师
	要	去	去	爱	想	妈妈
	来	来	要	看	同学	要
	眼睛	眼睛	来	去	能	想
	时候	想	同学	来	长	同学
	做	同学	时候	要	叫	能
	同学	时候	想	能	母亲	说
	想	能	能	同学	起	长
	能	做	喜欢	时候	班	叫
	喜欢	母亲	叫	想	头发	会
	走	喜欢	做	喜欢	知道	母亲
	叫	朋友	长	长	父亲	起
	起来	知道	母亲	叫	写	班
	起	字	朋友	做	带	爸爸
	知道	叫	班	母亲	没有	头发

从表中可以发现,每个位置的句子所使用的关键词有一些区分,但仍然不明显,仅仅依靠关键词对句子分类是不够的,实验很可能无法通过关键词来达到判别句子在段落中所处位置的目的。因此,考虑段落内部的句间的顺序是否能够作为判断逻辑合理的依据,并依此提出任务二,从整体上对文本内容的逻辑进行判别。

另外,从 S2PM 任务的实验结果中可以看出,SVM、朴素贝叶斯在分类效果中表现优于深度学习的 RNN 与 CNN。分析如下:实验中,SVM、Bayes 使用的均为由 fastText 模型得到的词向量文件,其对句子表示为句内的关键词词向量的加和求取平均数;而一般情况下,句子表示由句子内关键词的字典 id 拼接而得到,因此还实现了传统 SVM 以及 Bayes 使用关键词的字典 id 对句子进行表示的方式,对同一批数据进行了同样的分类实验。图 4 展示了传统词语表示方法与词向量分布式表示方法对于 SVM 以及 Bayes 分类的效果。

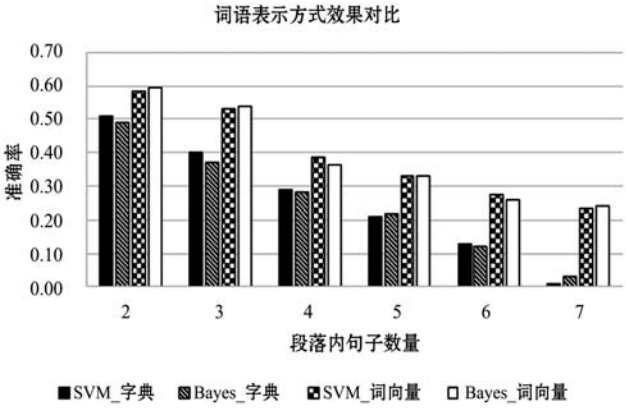

图 4　词语表示方法对于传统算法判别效果对比

从图 4 可知,传统 SVM 与 Bayes 仅仅依靠关键词完成 S2PM 的分类任务,分类效果差;而 S2PM 的实验中使用词向量,在考虑了关键词的情况下,将部分语义信息也融合在了分类任务中,语义信息的完善能够显著提高分类的效果,致使其在处理 S2PM 的分类任务中表现稍优于深度学习的 RNN 与 CNN。

3.4　DPM 实验效果及分析

由对 S2PM 实验结果的分析,关键词在各个位置的句子中的区分度并不大,依靠句子-位置模型判别段落逻辑是否合理,效果相对较差,因此,本节考虑对段落整体进行逻辑合理性判别。合理的段落结构,其句子的顺序是一定的,若其内部的句子顺序改变,几乎一定会影响到文本整体的连贯性。本节拟将整个段落作为输入,仍仅保留句中的动词、名词和形容词作为句子的特征,若输入的句子顺序正常,则对应标签为合理;若输入的句子顺序混乱,则对应标签为不合理,构建了段落整体逻辑合理性判别模型 DPM。实验结果如图 5 所示。

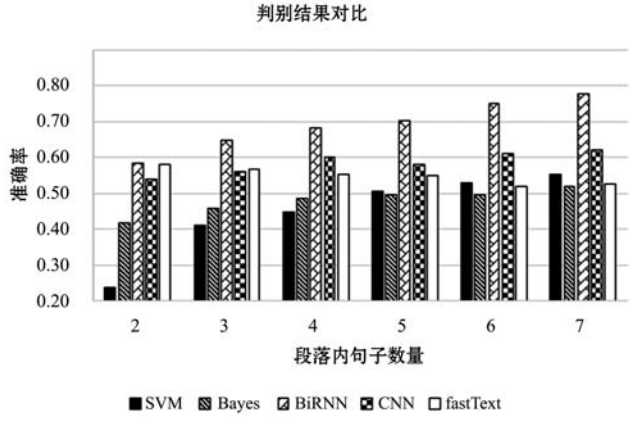

图 5 段落整体逻辑合理性判别结果对比

从图 5 中可以看出,BiRNN 在段落整体逻辑合理性的判别实验中表现最优,平均准确率达到 70% 左右,最优能够达到 78%。相比其他方法,BiRNN 更能够捕获前后句间的关联信息,这种优势源自其自身网络结构的设计,能够同时关注上下文捕获更多的信息。

3.5 DPM 优化特征的实验效果及分析

本文认为句子之间的连接,主要依靠前后句子相连接部分的关键词,BiRNN 能够更好地记忆句子连接信息,因此提出抽取当前句的句首与上一句的句尾的关键词表示句子信息的优化特征方法,这将能更好地提高分类器的分类效果[22]。将段落内的每个句子经过词性过滤之后,重新进行了特征选取,all 表示使用原始文本,e 表示使用每句话的末尾 N 个词语,b 表示使用每句话起始的 N 个词语。通过分析语料中每句话的词语数量,本实验设定 $N=2$。实验结果如图 6 所示。

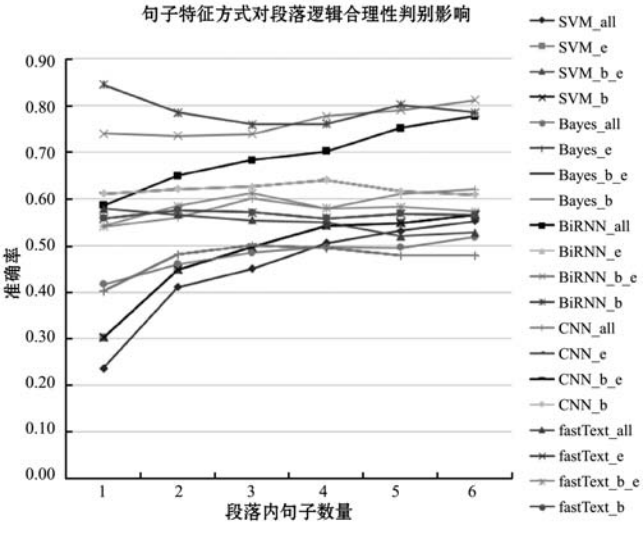

图 6 句子特征表示方式对判别效果影响对比

从图 6 中可以看出,在段落内句子数量在 3 ~ 6 之间时,使用新的特征方式,均能在一定程度上提升模型对段落逻辑合理性的判别准确率。在其他情况下,仅导致 Bayes 算法的判别准确率下降,其他算法的准确率均有所提升。综合考虑,模型应用 BiRNN 算法,保留句首或句尾关键词作为句子特征的方式,显著提高了准确率,准确率最高为 84%,且对每类段落的评判准确率均超过 75%。试验结果表明,句子中间部分单词在段落逻辑合理性判别的任务中噪声较大,使用全部关键词对分类器造成了一定的不良影响,因而对句子特征的进一步筛选,能够较好地保留句子携带的信息,达到优化去噪的效果。

4 结 语

本文研究的是中小学作文领域,从段落逻辑合理性的定性分析角度考量学生的语言运用能力,该研究具有现实意义。本文使用机器学习算法,构建的评测模型不需分析待测段落的上下文语义,即可进行逻辑合理性等级的自动评测,使用便利:构建句子-位置模型 S2PM 以及段落整体逻辑合理性判别模型 DPM,对段落逻辑合理性进行定性分析。将分类模型 SVM,Bayes、BiRNN、CNN、fastText 应用在对段落内句子位置、段落整体逻辑合理性判别上,分析其表现效果。实验结果显示,由于 S2PM 的构建依据段落内不同位置的句子内关键词区分度,卡方检验判定不同位置的句子所包含的关键词区分度并不大,导致 S2PM 取得的最好结果仍不令人满意。相比之下,DPM 中五种模型在对段落逻辑合理性判别的任务中的准确率更高,且以 BiRNN 最有效。通过对句子特征-关键词的进一步筛选,句首关键词、句尾关键词或两者结合共同作为句子特征,比句子全部关键词更能代表句子信息,由此可以推断出句子中间部分单词在对段落逻辑合理性判别的任务中噪声较大。精简句子关键词,在一定程度上减少了语料中存在的噪声。同时,试验结果也表明了使用分类模型对段落逻辑合理性判别是有效的,为以后对文本逻辑合理性判别算法及其应用做进一步研究和改进提供了参考。

参 考 文 献

[1] Ponte J M, Croft W B. Text segmentation by topic[C]//European Conference on Research and Advanced Technology for Digital Libraries. Springer-Verlag, 1997:113 - 125.

[2] 傅间莲,陈群秀. 自动文摘系统中的主题划分问题研究[J]. 中文信息学报,2005,19(6):28 - 35.

[3] 林林. 基于规则的文本过滤系统设计与实现[D]. 广州:华南理工大学,2003.

[4] 郑婧,孙卫. 国内自然语言处理技术研究与应用的状态[J]. 数字图书馆论坛,2008(7):27 - 31.

[5] 黄曾阳. HNC 理论与自然语言语句的理解[J]. 中国基础科学,1999(S1):85-90.
[6] 黄曾阳. HNC(概念层次网络)理论:计算机理解语言研究的新思路[M]. 北京:清华大学出版社,1998.
[7] 司贝贝,杨进才. 基于依存关系的复句关系词搭配库建设[J]. Software Engineering & sapplications,2015,4(4):81-87.
[8] 李艳翠,孙静,周国栋. 汉语篇章连接词识别与分类[J]. 北京大学学报(自然科学版),2015,51(2):307-314.
[9] 杨进才,郭凯凯,沈显君,等. 基于贝叶斯模型的复句关系词自动识别与规则挖掘[J]. 计算机科学,2015,42(7):291-294.
[10] 刘挺,车万翔,李正华. 语言技术平台[J]. 中文信息学报,2011,25(6):53-62.
[11] Xue N. Chinese word segmentation as character tagging[J]. 中文计算语言学期刊,2003, 8(1):29-47.
[12] 杨宪泽. 21 世纪高校特色教材,人工智能与机器翻译[M]. 成都:西南交通大学出版社,2006.
[13] Zhang H. The optimality of naive bayes. [C]//Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, Usa. 2004.
[14] Schwenker F. Hierarchical support vector machines for multi-class pattern recognition[C]//International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. Proceedings. IEEE, 2000:561-565 vol.2.
[15] Osuna E, Freund R, Girosi F. Training support vector machines: an application to face detection[C]//Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 1997:130-136.
[16] 忻栋,杨莹春,吴朝晖. 基于 SVM—HMM 混合模型的说话人确认[J]. 计算机辅助设计与图形学学报,2002, 14(11):1080-1082.
[17] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning [C]//International Joint Conference on Artificial Intelligence. AAAI Press, 2016:2873-2879.
[18] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[EB]. eprint arXiv:1404.2188v1, 2014.
[19] Kim Y. Convolutional neural networks for sentence classification[EB]. eprint arXiv:1408.5882, 2014.
[20] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification [J]. eprint arXiv: 1607.01759, 2016:427-431.
[21] Shamma D A, Shamma D A, Friedland G, et al. YFCC100M: the new data in multimedia research[J]. Communications of the Acm, 2016, 59(2):64-73.
[22] Luhn H P. The automatic creation of literature abstracts [M]. IBM Journal of Research and Development, 1958, 2(2):159-165.

(上接第 53 页)

[7] Youman C E, Sandhu R S, Feinstein H L, et al. Role based access control models[J]. Information Security Technical Report, 1996, 6(2):21-29.
[8] National Institute of Standards and Technology. Federal Information Processing Standard(FIPS). US, Department of Commerce, Advanced Encryption Standard[OL]. 2001. http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf.
[9] 闫乐乐,李辉. 基于复合混沌序列的动态密钥 AES 加密算法[J]. 计算机科学,2017,44(6):133-138,160.
[10] 张险峰, 秦志光, 刘锦德. 椭圆曲线加密体制的性能分析[J]. 电子科技大学学报, 2001, 30(2):144-147.
[11] Gentry C. Fully homomorphic encryption using ideal lattices [C]//Proceedings of the Annual ACM Symposium on Theory of Computing, 2009:169-178.
[12] 韩天悦,谢静. RSA 加密解密算法及相关攻击方法[J]. 电脑与信息技术,2018,26(1):53-55.
[13] 缪昌照,徐俊武. AES 与 ECC 混合加密算法研究[J]. 软件导刊,2016,15(11):63-64.
[14] Sweeney L. k-Anonymity: A Model for Protecting Privacy [J]. International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems, 2002, 10(5):557-570.
[15] Dwork C. Differential Privacy: A Survey of Results[C]//International Conference on Theory and Applications of MODELS of Computation. Springer-Verlag, 2008:1-19.
[16] 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1):101-122.
[17] Machanavajjhala A, He X, Hay M. Differential Privacy in the Wild: A Tutorial on Current Practices & Open Challenges[J]. Proceedings of the Vldb Endowment, 2016, 9(13):1611-1614.
[18] Zhang J, Xiao X, Xie X. PrivTree: A differentially private algorithm for hierarchical decompositions[C]//Proceedings of the 2016 International Conference on Management of Data. ACM,2016:155-170.
[19] 朱作玉. 支持隐私保护的极限学习机研究[D]. 沈阳:东北大学, 2014.
[20] 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014(4):927-949.
[21] Zhang M, Deng Z, Che W, et al. Combining Statistical Model and Dictionary for Domain Adaption of Chinese Word Segmentation[J]. Journal of Chinese Information Processing, 2012, 26(2):8-12.