

眼动记录与主旨结构标注的关联性分析研究

单昊聪, 周强

清华大学信息技术研究院

语音和语言技术中心, 北京, 100084

2390326004@qq.com, zq-lxd@mail.tsinghua.edu.cn

摘要

给定包含主旨概括句的汉语句群, 针对该句群的内部结构标注是基于语言学的分析结果, 而阅读句群时的眼动轨迹则蕴含着人的心理认知, 两者的信息融合和内在关联性分析是该文主要工作。该文使用基于径向基函数支持向量机和递归特征消除的分类模型, 根据标点小句片段对应的眼动指标数据预测该片段是否为包含主旨内容的关键信息, 达到了0.76的准确率, 并通过分析关键片段上眼动数据的分布特点, 提取出对句群主旨概括信息区分度较好的眼动指标。

关键词: 眼动指标; 文本结构标注; 支持向量机; 径向基函数; 特征递归消除

Research on the correlation between eye movement feature and thematic structure label

Haocong Shan, Qiang Zhou

Center for Speech and Language Technology,

Research Institute of Information Technology,

Tsinghua University, Beijing, 100084, China.

2390326004@qq.com, zq-lxd@mail.tsinghua.edu.cn

Abstract

Given a Chinese sentence group that contains a theme sentence, the internal structure label of the sentence group is based on the results of linguistic analysis, while the eye movement trace of reading sentence group contains human psychological cognition. The main work of this paper is the information fusion and internal relevance analysis of structure label and eye movement. In this paper, a classification model based on radial basis function support vector machine and recursive feature elimination is used to predict whether the punctuation clause segment is the key information containing the thematic content according to the corresponding eye movement feature data. The accuracy of 0.76 is achieved. By analyzing the distribution characteristics of eye movement data on the key segment, eye movement features with good distinction for the thematic information of the sentence group are extracted.

Keywords: Eye Movement Feature, Text Structure Label, Support Vector Machine, Radial Basis Function, Feature Recursion Elimination

本作品已根据《Creative Commons Attribution 4.0 International Licence》获得许可。许可证详细信息: <http://creativecommons.org/licenses/by/4.0/>.

1 引言

人类使用眼睛运动来选择性地获取视觉信息再加工处理是一个很重要的认知过程。在人类阅读相关的心理学研究中,挖掘眼动数据所蕴含的信息是重要的方向。一方面,眼动数据包含着读者阅读行为的认知信息。例如不同教育背景读者的阅读能力的差异性可以反应在眼动模式中,更详细的研究显示不同读者的阅读策略等信息也可以依据眼动数据来判别 (Ablinger et al., 2014)。另一方面,阅读任务的不同也影响着眼动模式。例如无意识阅读和有目的阅读的眼动指标分布有着较明显的差异 (Reichle et al., 2010)。针对有目的阅读而言,不同阅读任务所对应的眼动模式也是不同的。例如在文本校对和文本理解两个不同阅读任务中,眼动数据所反应的注意力分配是不同的 (Kaakinen and Hyönä, 2010)。在文本主旨相关的任务中,有融合眼动数据的多主旨文本标题提取任务研究 (Hyönä and Lorch, 2004),该研究主要考察了标题句的眼动数据,并详细分析了眼动指标的特性,但是该研究所选取的文本缺少具有语言学支撑的结构标注,同时缺少衡量眼动指标的重要性分布。也有单主旨文本摘要提取的眼动数据研究 (Xie et al., 2019),但是该研究仅仅专注于眼动轨迹,没有完整的眼动指标分析,缺乏认知心理学依据。

在分析阅读的眼动研究中,主要包含注视和眼跳两种基本的眼动现象 (闫国利 et al., 2013),第一类是与眼睛何时移动有关的时间维度眼动指标,例如遍注视时间,其反映了读者处理特定信息所需要的时间;另一类指标是与眼睛移动位置有关的空间维度的眼动指标,具体包括眼跳距离,注视位置,注视顺序等,其中注视位置反映了读者在特定时间内所处理的信息内容在上下文中的位置,注视顺序则反映了读者处理信息的次序。眼动数据在更深层次上还反映了读者在词汇和句子加工理解过程中的复杂心理认知阶段。Rayner (1979)在研究中使用了针对不同词汇的“首次加工有多次注视的首次注视时间”这一指标,较好地反映了读者对于词汇的早期加工情况,同样地第二次注视时间也被看做是一个较好反应复合词早期加工的指标 (Pollatsek et al., 1986)。第二遍阅读时间和回视时间等则反映了读者对于复杂句子中词语的后期处理情况。后期处理的指标往往反映了被试在兴趣区(包含字、词或句子等的区域)遇到困难后的再处理和分析过程。所以时空维度的眼动指标数据可以反应人在阅读语篇时候的心理认知过程。这样的心理认知过程蕴含着人对于语篇中词语的辨识、句子的解读和整个语篇的理解。

该文旨在融合基于语言学的主旨结构标注和基于认知心理学的眼动数据指标信息,建立模型来探索两者的关联性,并考察眼动指标的显著性分布,提取出在主旨判断中重要的指标并做心理学分析,以填补目前的研究空白。该文主要面临三个问题。首先是眼动数据来源于人对文本的阅读,选择合适的文本供被试阅读是很重要的实验前提,需要从语言学的角度将文本做处理,以能够配合眼动数据做可计算的数据集。其次是如何根据数据集选择合适的计算模型,在保证计算模型的精度情况下,需要挖掘出眼动指标数据对于语篇分析的指导意义,提取出较为重要的眼动指标组,从认知角度给予语篇意义分析一定的帮助。最后是如何设计实验,需要考虑在研究历史中眼动数据的特点合理地设计实验以保证实验的科学性。

对于文本的选择问题,该文的文本研究对象为句群 (吴为章, 2000)。较广泛意义上的语篇而言,句群在语义上有逻辑联系,在语法上有结构关系,一般都会会有一个明晰的中心意思即主旨。主旨是句群全部内容表达的基本观点,是其叙写议论的基本描述目标。按照其归纳形式不同,大致可以分为显性表示和隐性表示两类。在显性表示类中,句群主旨直接落在句群的主题句中,可以通过对主题句内容的直接引用或适当改写得到。在隐性表示类中,句群中不存在明显的主题句,但大多存在包含句群关键信息的核心句,通过对这些核心句子内容的概括提炼,可以生成合适的主旨内容描述句。在连贯话语中,句群是相对独立的句法语义结合体,可以从语流中切分出来。在语篇结构分析标注研究中,它起着承上启下的作用。按照其中描述内容不同,大致可以分出记叙、议论、说明等功能类别,它们具有不同的结构描述模式。对于这样的文本,其含有的信息价值较高,眼动数据的可研究性也更为丰富,并且能够和文本摘要任务很好的契合,因为能够依据语言学研究对文本做主旨结构标注,以产生可以为模型训练的数据。所以此时就确定了研究的主要范围为汉语句群片段的眼动数据和主旨标注数据之间的信息融合和关联性。针对阅读汉语句群片段,一方面通过被试阅读该汉语句群,记录下眼动数据,另一方面通过基于语言学的句群主旨结构分析,可以得到人工标注数据,这两部分的数据结构细节在第2节数据介绍中做详细的说明。

对于模型的选择问题。在获得了数据集后,需要针对此数据集合理地选择模型做训练和预

测。随着统计机器学习方法和神经网络的发展 (Lauzon, 2012), 以及数据量越来越大的真实世界数据现象, 机器学习方法越来越广泛地应用到实际数据分析中。虽然神经网络的精确率和性能要比统计学习方法更好, 但是由于统计学习方法有着更为深厚和坚实的数学理论基础, 所以在可解释性 (Molnar, 2020) 方面统计学习方法更胜一筹。该文研究的重要方向是眼动数据指标的描述特点, 分析其对于文本主旨标注的关联性, 在保证模型精度的同时, 还需要对关联性有较好的解释, 所以对于模型而言选择了统计机器学习方法中较为经典的支持向量机 (Suykens and Vandewalle, 1999) 模型, 对更为详细的模型设计在第3节给出。

对于实验的设计问题。在获取了数据集并确定了模型后, 需要考虑影响实验的因素来合理地设计。首先, 对于人的阅读认知过程而言, 不同的人有着不同的阅读理解水平, 并且该阅读水平也反映在了眼动数据的模式上 (Hyönä et al., 2002)。所以为了减少眼动数据的误差, 需要考虑不同背景的被试对于结果的影响, 并提取出较优的被试集合来达到更高的准确率。其次, 考虑到在人工文本主旨标注过程中, 有一个环节是要求总结出句群的主旨信息并添加到标注数据表格中, 这一环节可以看作是带主旨总结目的的阅读过程。普通的自由阅读和带主旨总结目的阅读的心理认知是不同的 (Kaakinen and Hyönä, 2010), 其差异性也会反应在眼动数据中, 所以需要进行对比实验, 即自由阅读和带主旨总结任务阅读两个阶段的眼动实验。通过对比来分析指标的差异性。实验阶段的详细信息和实验结果在第4节给出。

2 数据介绍

该文的原生数据分为两部分, 一部分是微博和百科类句群的人工标注数据, 另一部分是对每个语篇的眼动数据。在获取了两部分数据后, 将处理后的数据做匹配得到可训练的数据集。

2.1 微博和百科类人工标注数据

对于微博和百科类的语篇而言, 对其进行结构主旨标注 (周强, 2019) 来分析挖掘其中主旨信息的不同表现模式。

首先根据其中的各个结句点号 (句号、问号、叹号、省略号等) 自动切分出各个句子 (Sentence), 在这个过程中, 需要注意标点 (双引号、括号等) 内部的分隔作用。然后在各个句子中进一步依据其中的分隔点号 (主要是逗号、分号和冒号) 自动切分出各个标点小句片段 (Punctuated Clause Segment, PCS), 至此PCS层小句产生。

下面表 1 给出一个具体的微博短讯标注实例对此进行解释说明:

句子序号	句子和小句	主旨描述
1	前天晚上, 武昌警方调集数百名警力, 对武汉大学校园周边治安环境进行整治。	武昌警方在武大校园周边查验8名涉嫌吸毒男子。
1-1	前天晚上,	
1-2	武昌警方调集数百名警力,	
1-3	对武汉大学校园周边治安环境进行整治。	
2	在记者跟随下, 一队民警从武大近邻的军悦假日酒店内揪出8名涉嫌吸食麻果的男子。	
2-1	在记者跟随下,	
2-2	一队民警从武大近邻的军悦假日酒店内揪出8名涉嫌吸食麻果的男子。	
3	他们都坚决否认吸毒, 民警当面查验, 证实其中6人吸了毒。	
3-1	他们都坚决否认吸毒,	
3-2	民警当面查验,	
3-3	证实其中6人吸了毒。	

表 1: PCS层切分实例: 微博短讯 (武汉晚报-20121020-1600)

在表 1 中, 可以看到该武汉晚报微博短讯里包含有三个句子, 每个句子中又包含有多个PCS层小句片段, 例如在句子序号为1的句子中, 包含有1-1, 1-2和1-3三个PCS层小句。主

旨描述列信息则是从句子形成的句群中抽取关键信息总结得到，此时完成了对整个语篇基于PCS层细粒度的划分。然后需要进一步确定PCS层小句为关键或非关键信息。关键信息的标注则可由结构主旨标注相关信息自动得出。

句子序号	句子结构序列	小句结构序列	关键信息
1	ES		
1-1		TIM	0
1-2		FE-1	1
1-3		FE-1	1
2	ES		
2-1		BE-2	0
2-2		FE	1
3	OS		
3-1		BE	0
3-2		FE-1	0
3-3		FE-1	0

表 2: 实例所对应的结构标注信息

在表 2 中，展示了与表 1 中 PCS 层一一对应的主旨结构标注信息，关键信息列中的标注信息需要直接用于实验。其中标注为 1 代表关键信息，标注为 0 代表非关键信息。而关键信息列的标注信息由句子结构序列和小句结构序列的标注信息而得到。例如句子 1 和 2 中，句子结构序列标注为 ES (Event Sentence) 事件句，表示其包含句群的部分关键信息。进一步考察其中的小句结构序列标注，可发现 FE 代表前景小句，包含了句子的主要事件描述内容，所以相应的 PCS 小句的关键信息标注为 1。而句子 3 的句子结构序列标注为 OS (Other Sentence)，不包含句群关键信息，只是对关键句的补充解释或者阐释说明，所以其后的相应 PCS 的关键信息标注为 0。

利用上述方法对 1172 个微博短讯和 1300 个汉语新闻百科句群的人工标注数据进行自动提取，得到可用于实验的标注数据，以匹配眼动数据做为实验数据集。

2.2 语篇眼动数据

眼动实验一共有 125 位被试者，每位被试者要求分两个阶段使用眼动仪阅读 2.1 中的微博和百科语篇。一个阶段是正常自由阅读语篇，简称 *article* 阶段；另一个阶段需要被试者在阅读完毕后总结语篇的主旨，简称为 *gist* 阶段。同时每个语篇需要至少三位以上的被试做实验，可以获得至少 3 份不同的被试记录数据，以为后面的数据处理减少误差做准备。

实验通过眼动仪产生注视眼动记录报告，所有不同被试的相关数据都会保存在一个 EXCEL 表中，可以根据 *Eyelink* 用户手册描述自动导出不同被试的相关数据报告 (Research, 2010)。报告中较为重要的是兴趣区数据和眼动指标数据。该文兴趣区设置为句群中的 PCS 层次小句片段，所以实验中的兴趣区即为 PCS 层小句。针对眼动指标数据，由于存在冗余的指标，需要对部分指标做选取和舍弃。这涉及特征工程中的特征选择问题。在特征工程的数据预处理中，对于实际意义差距很大的特征数据一般采用独热编码 (Jiang et al., 2016)。而对于眼动注视数据而言，其特征大部分是与时间尺度相关的变量，其实际意义相近或相同，例如第一次注视时间等；另一部分是次数相关的指标，例如注视该兴趣区次数，此情形下采用 *z-scores* 标准化数据 (Abdi, 2007) 较为合理。

同时需要保证每个特征的数据在标准化后的数值具有相同的度量标准，那么对于个别眼动特征而言，例如开始时间、结束时间，其数值为绝对时间点而非可以衡量时间长短的区间值，所以此类特征也要舍弃。另外在实验中，被试和仪器会有短暂的数据未记录的现象，多数表现为数据为空值，如果按照上述的特征选择标准，结合特征的实际意义，这些缺少值可以直接赋予 0。例如若注视时间为 0，则代表被试未注视该兴趣区；若注视次数为 0，则同样代表被试为注视该兴趣区。对于分类模型而言，这也是减少误差的有效方法。

另一方面眼动特征主要有注视和眼跳两个方面。注视数据报告中每一行表示注视点的各种对应指标值，按照注视点的发生时间顺序排列。眼跳数据报告每行则表示一个眼跳的各种对应指标值，按照每个眼跳的发生时间顺序排列。注视数据和眼跳数据实际上属于眼动指标中时间

和空间两个不同维度的信息。该文以注视相关的眼动指标为考察重点，在未来的实验中会加入眼跳相关的数据分析。

经过上述过程就可以在PCS层次建立起眼动兴趣区注视相关指标。而同时句群数据的结构主旨标注也是基于PCS层次，这样两者就形成对应关系，以进行后续的信息融合和关联性研究。至此可以得到16个筛选出的眼动指标，相应的实际意义如表 3所示：

指标名称	含义
IA_DWELL_TIME	所有注视点总注视时间
IA_FIXATION_COUNT	该兴趣区的所有注视点个数
IA_RUN_COUNT	注视该兴趣区次数（注视点落入后离开算注视一次）
IA_LAST_FIXATION_RUN	该兴趣区阅读遍数
IA_FIRST_RUN_DWELL_TIME	第一遍进入兴趣区到离开的所有注视点时间之和
IA_FIRST_RUN_FIXATION_COUNT	第一遍阅读这个兴趣区的注视点个数
IA_FIRST_FIXATION_DURATION	第一个注视点的注视时间
IA_SECOND_RUN_DWELL_TIME	从第二次进入兴趣区到离开所有注视点时间之和
IA_SECOND_RUN_FIXATION_COUNT	第二遍阅读这个兴趣区的注视点个数
IA_SECOND_FIXATION_DURATION	第二个注视点的注视时间
IA_THIRD_RUN_DWELL_TIME	从第三次进入兴趣区到离开所有注视点时间之和
IA_THIRD_RUN_FIXATION_COUNT	第三遍阅读这个兴趣区的注视点个数
IA_THIRD_FIXATION_DURATION	第三个注视点的注视时间
IA_LAST_RUN_DWELL_TIME	最后一次进入兴趣区的总时间
IA_LAST_RUN_FIXATION_COUNT	最后一次进入兴趣区的注视次数
IA_LAST_FIXATION_DURATION	最后一个注视点的注视时间

表 3: 眼动指标说明

3 模型设计

该文基于被试眼动数据判断所对应的PCS层小句是否为关键信息句，来挖掘不同的眼动指标的重要性排序以及相应的心理学依据。根据第2节的实验数据描述可以清楚地看到需要训练一个二值分类器，依据此分类器对相应特征的重要性做排序。

另外，还需要考虑一个很重要的数据特征，即数据是线性可分还是非线性可分。在2.2的表 3中可以看到，眼动指标比较复杂。这些指标大致分为四个主要类型：第一类指标为注视点总特征，例如所有注视点总注视时间以及个数；第二类指标为第一遍进入兴趣区相关的注视时间和注视点个数；第三类指标为第二、三遍阅读兴趣区的注视点信息；第四类为最后一次进入兴趣区的总时间。实际上这些指标之间可能有着较为复杂的信息交叉。例如所有注视点的总注视时间和所有注视点的个数可能会成正相关。因为在普通阅读中，可能存在着注视点个数越多，被试看的时间越久的情况，实际上该文计算了这两个指标的相关系数 (Lee Rodgers and Nicewander, 1988)为0.35，所以其并不是完全的正相关。可能出现注视点个数较少，但是每个注视点的时间较久导致总注视时间大的情况。尽管如此这两个指标之间的信息还是存在一些交叉，其他几类的指标仍然可能存在着类似的情况。所以在实验的过程中需要考虑数据是否是线性可分的，如果非线性可分，那么应该有线性和非线性的结果比较，所以此模型应该能同时处理线性可分数据和非线性可分数据。

另外对于特征的重要性衡量也是需要考察的很重要的方面。因为在认知心理学中，对于眼动指标的认知方向的研究较为丰富。如果通过眼动数据来判断该眼动数据对应的小句片段是否为关键信息，那么也应当分析该眼动数据是如何起作用的。例如眼动数据中哪一部分指标起着重要作用，哪一部分指标的作用较小，以此获得不同眼动指标的显著性分布。依据该分布可以提取出对于主旨结构信息比较重要的指标，并结合认知心理学做分析。所以模型需要具备衡量指标重要性的功能。

综上可知模型需要根据PCS层次上的眼动指标数据来判断该片段是否关键，无论这些数据是线性可分或非线性可分，同时还需要对这些眼动指标如何作用的过程有着较为清晰的展现，能够可视化指标的重要性排序以及显著性分布。结合这些模型需求和数据特点，该文选择支持

向量机SVM(Support Vector Machine)和递归特征消除RFE(Recursive Feature Elimination)相结合的模型: SVM-RFE。因为支持向量机本身是一个数学理论基础较为完善的模型, 同时对于近似线性可分的数据有较好的处理结果, 并且对于非线性可分数据而言, 可以通过核函数等技巧来解决。另外支持向量机结合递归特征排序可以对特征的显著性做分析。

实际上, SVM-RFE算法是一个经典的特征排序方法, 适用于数据是线性可分或者近似线性可分的情况。它使用SVM线性核, 并且通过RFE包装方法来将特征做筛选和排序。筛选排序的依据就是SVM训练过程的参数 w , w 反映了所有特征的权重。该权重在空间中反映了每个维度的特征对结果的影响情况, 所以可利用该权重的大小对特征重要性做排序。对于SVM-RFE算法 (Guyon et al., 2002)本身, 其正确性和可靠性已经在分类患者和读者等实验中有了很好的验证: 特征权重的大小一定程度地反映了该特征对于模型预测结果的重要程度, 对于结果没有较大影响的特征可以做筛减。

至此, 该文确定采用SVM-RFE算法做训练和分类。在实验的结果中发现, 对于注视眼动指标而言, 线性SVM不能很好地区分出关键句和非关键句。在下文第4节结果分析里以及上述的数据分析中也可以推测眼动指标数据并非是非线性可分的, 所以线性分类模型不适用于注视眼动类数据。

对于非线性可分数据而言, SVM中的非线性核, 例如径向基核函数RBF(Radial Basis Function) (Musavi et al., 1992), 可以提供一种处理此类型数据的方法。该核技巧的思想是将原数据的有限特征映射到更高维度的特征空间中, 使得数据在更高维度的特征空间中能线性可分。

但是该文需要研究不同眼动特征对于主旨判断的认知意义, 而非线性RBF核函数的主要缺陷是没有线性核函数中的特征权重 w 。该方法专注于训练后的精确度结果, 对于映射后的高纬度特征空间的函数不能够显式地表达, 其背后原因是将RBF核函数做泰勒展开后, 得到的是无穷维的特征映射函数 (李航, 2019)。所以该方法不能做特征筛选和排序。

针对这个缺陷, 很多研究提供了改进办法。例如有的研究使用高斯核函数, 专注于改进核函数的参数选择难题, 并结合RFE方法来抽取出非线性特征 (Xue et al., 2018)。有的研究引入对角矩阵, 将特征参数放置于矩阵对角线中, 其余的矩阵位置归零, 并通过迭代的方式将映射后的特征不断削减, 最终得到有限个特征的排序 (Mangasarian and Kou, 2007)。这些研究进展大多不需要考察指标的实际意义, 只需要专注于准确率的提升。但是对于该文而言, 指标的重要性分布以及其在心理认知中的含义是考察的重点。

该文使用了对于RBF核函数做变形展开的处理思想 (Liu et al., 2011), 将核函数的表达式改写为内积形式, 同时改写SVM的目标优化函数, 将核函数和目标函数结合成包含内积形式核函数的SVM目标优化函数。然后对此函数做麦克劳林展开, 最后对隐函数求偏导, 可以得到一个计算特征权重的表达式。此类表达式的处理类似于线性核函数中的参数 w , 所以非线性RBF核函数的特征权重可以计算。另外由于非线性RBF在训练的过程中有超参数, 该文在训练过程中加入了交叉验证方法以得到最佳的参数值。

4 实验结果分析

实验方法是把所有的PCS层小句对应的眼动数据以及相应的关键标注信息作为输入, 用3节中的算法处理训练模型来判断该小句是否为关键信息, 同时衡量眼动指标的显著性分布。实验主要考虑三个变量, 首先是线性和非线性SVM对于实验结果的影响, 其次是不同被试以及不同阅读任务的影响, 后两个变量对于认知心理学判断有着重要的意义。

在认知心理学中, 不同受教育程度的被试有着不同的阅读认知水平, 不同被试的受教育程度影响着对文章意义把握的准确率 (Lou et al., 2017), 这些信息蕴含在眼动数据中。所以需要考察不同被试对于实验准确率的影响。另一个重要的变量是不同的阅读阶段, 在2节中可看到被试参与两个不同的阅读阶段, 自由阅读简称为*article*阶段, 和总结主旨简称为*gist*阶段, 需要考察这两个阶段的阅读特性和眼动特征。

该文根据这三个因素设计了三个实验。实验一侧重于线性和非线性SVM-RFE的比较分析; 实验二建立在实验一的结果之上, 在确定使用非线性SVM-RFE模型后, 考察不同被试的影响, 选取一组质量最好的被试作为数据集; 实验三建立在实验二的结果之上, 在选定了质量较好的被试作为数据集后, 着重*article*和*gist*阶段的结果分析。

4.1 实验一结果分析

对于实验一而言，训练了线性和非线性SVM两种模型，其中线性SVM中惩罚参数 c 为1，非线性RBF核函数中 γ 为0.1，均采用交叉验证方法得到。两种模型的结果如下所示：

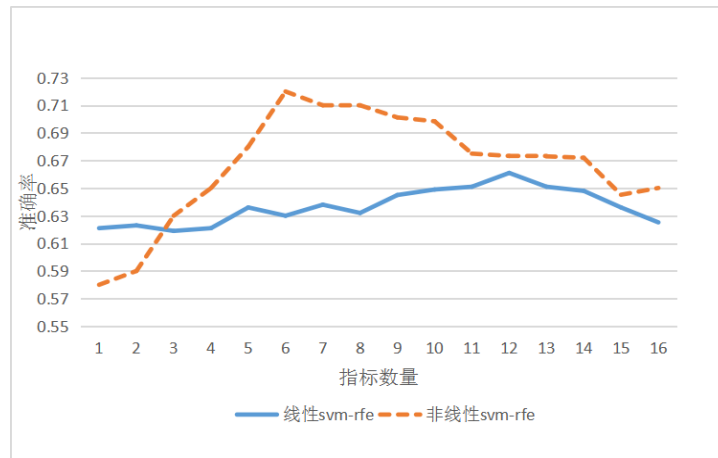


图 1: 线性和非线性比较

在图 1中可以看到对于线性SVM而言，不同的指标特征数量对准确率的影响很小，总体在0.65左右波动，表明线性SVM并不能够很好地将数据分割开。数据在特征空间很大程度是非线性可分的，可能因为眼动指标之间较为复杂，指标和指标之间的有一定的关联性，并不是所有的指标都具有非互性。

而对于非线性SVM-RFE而言，一方面可以看到其准确率随着指标的选择数有明显的变化，能够反应出不同指标对于准确率的影响程度，另一方面，在指标数通过RFE筛减到6的时候，准确率达到最高，说明前6个指标的综合作用效果最为显著。这也验证了非线性SVM-RFE对于眼动数据指标分类的可行性。

4.2 实验二结果分析

在实验一的基础上，可以得到非线性优于线性分类器的结论，此时可以确定实验二的基本模型为非线性SVM-RFE分类器。实验二需要考虑不同被试对于实验的影响，因为在选择被试的时候并不能够具体确定其阅读水平。

该文按照125位被试切分数据集，每个被试形成一个独立的数据集，该数据集内包含若干个语篇的眼动数据。将每位被试的数据集分别用非线性SVM-RFE分类器训练。得益于SVM方法对小样本数据训练的可靠性，虽然切分后的数据量小于切分前，但是结果依然稳定。被试准确率分布情况如图 2所示：

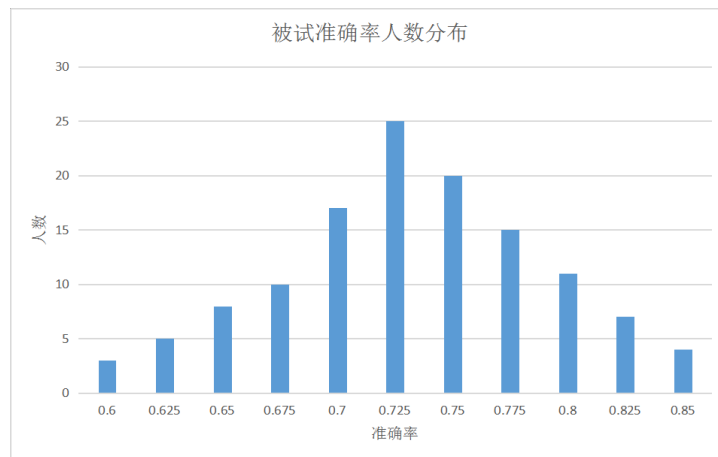


图 2: 被试准确率人数分布图

可以看到不同被试的准确率有着一定的差异性。整体而言，所有被试的准确率分布近似于正态分布，在0.725-0.75区间内的被试人数最多。通过该结果可以得到被试准确率的排序，准确率较低的被试可能会成为实验的误差。考虑到每篇语篇都有至少三位被试参与了实验，所以可以依据被试的准确率排名从高往低筛选语篇眼动数据，能够保证每篇语篇的眼动数据都是最优，最终可以得到一个整体眼动数据最优的数据集。

4.3 实验三结果分析

在实验一和二中，确定了非线性RBF核方法和基于区分被试的最优数据集。实验三就使用该最优数据集训练基于非线性RBF核的SVM—RFE模型。要注意的是在实验三中，数据集依据被试阅读的两个阶段划分为两个部分并分开训练。这样就可以做*article*阶段和*gist*阶段的对比分析。

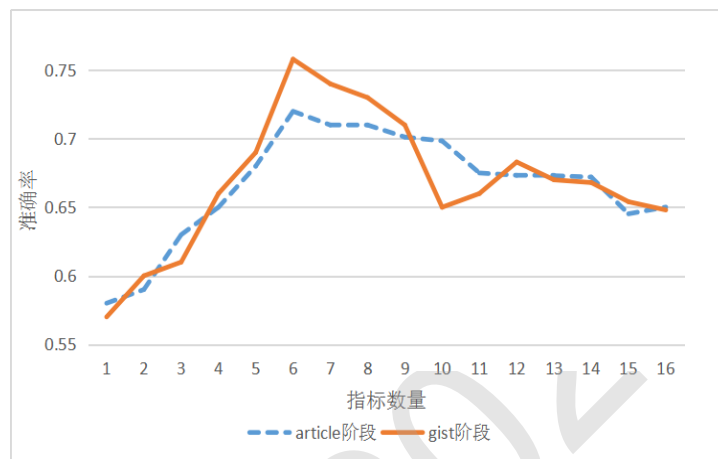


图 3: 不同阶段比较分析

图 3给出了两个不同阶段的准确率折线图。可以看到对于整体趋势而言，*gist*阶段和*article*阶段较为相似，都是在指标数为6个左右时准确率达到最佳。但是*gist*阶段的最高准确率为0.758，相应地比*article*阶段的准确率高一些。这也反映出*gist*阶段的被试数据质量要更高，即带有概括主旨任务的被试数据更具备判断关键信息的模式。

表 4给出了*article*和*gist*两个阶段的眼动特征重要性排序表，最左列为指标名称以及部分含义解释，右侧两列为两个不同阶段的指标重要性排序，数字越小表示重要性越高。表中*article*列按照重要性从高到低降序排序，*gist*列为与*article*相对应的排序。这样处理方法可以更清楚地对比两个阶段的异同。

整体而言，*article*和*gist*两个阶段眼动指标的整体排序有相似之处也有不同之处。首先，对于排名前6的最为重要的指标而言，*article*和*gist*都包含了所有注视点总注视时间、该兴趣区所有注视点个数、注视该兴趣区次数和该兴趣区阅读遍数四个指标。这四个指标反应的是被试对于一个PCS层小句的阅读频率和阅读时长。为了判断小句是否为关键句，被试对关键信息的阅读频率会更高，甚至多次多遍阅读。而对于非关键信息而言，正好相反。该结果可以从两方面来解读。一方面，总注视点个数以及次数在心理学里反映了被试对于阅读材料的认知加工负荷情况 (Henderson and Ferreira, 1990)，认知负荷较大的材料，注视点个数以及次数也越多。而含有主旨信息的内容可以被看作是句群中需要重点加工处理的对象，所以在主旨判断过程中，该四个指标的重要程度靠前是合理的。另一方面，不同阅读水平的被试阅读同一材料的注视点总个数和次数也不同，熟练阅读者的次数要明显少于不熟练阅读者的次数 (Rayner et al., 2011)，而实验二中的筛选被试环节也减少了读者熟练度差异所导致的误差。所以这四个指标的重要性正确地反映了主旨提取认知过程，同时被试的筛选环节也使得结果更加可靠。

其次，*article*和*gist*阶段指标有着不同的分布特性。在*article*阶段，含有总时间和次数的特征重要性靠前，第一遍阅读相关的特征其次，第二遍、第三遍阅读的相关特征重要性位于最后。而在*gist*阶段，最后一遍的阅读时间次数特征重要性靠前，含有总时间和次数的特征重要性其次，第一遍和第二遍特征重要性排在最后。这也符合人的阅读理解认知过程。在*article*阶段，被试实际上不会对关键信息重复确认，大部分句子都只会阅读一遍，如果第一遍阅读相关

指标含义	article排序	gist排序
所有注视点总注视时间	1	4
该兴趣区的所有注视点个数	2	3
注视该兴趣区次数 (注视点落入后离开算注视一次)	3	6
第一遍进入兴趣区到离开的所有注视点时间之和	4	11
该兴趣区阅读遍数	5	5
第一遍阅读这个兴趣区的注视点个数	6	8
最后一次进入兴趣区的注视次数	7	1
最后一次进入兴趣区的总时间	8	2
第三遍阅读这个兴趣区的注视点个数	9	7
从第三次进入兴趣区到离开所有注视点时间之和	10	10
第二遍阅读这个兴趣区的注视点个数	11	9
从第二次进入兴趣区到离开所有注视点时间之和	12	12
第一个注视点的注视时间	13	14
第三个注视点的注视时间	14	13
第二个注视点的注视时间	15	16
最后一个注视点的注视时间	16	15

表 4: 眼动指标两阶段排序

的时间和注视点个数较多, 说明被试在该小句上花费的时间较久, 则可认为此句包含较重要的事件信息的可能性更大。第一遍阅读属于前期加工指标, 反映了被试对于信息的预处理情况。这个结果也反映了被试在信息的预处理阶段就可以大致分辨出句群中的关键信息。在 *gist* 阶段, 被试需要在阅读后总结主旨, 此时最后一遍阅读相关文本的信息较为重要。在人的阅读行为中, 涉及较长文本的语篇时, 被试需要首先通读全文, 然后找到包含主旨信息的小句着重阅读, 提取主要信息。这些认知过程则直接反应在最后一遍阅读行为上, 在关键信息句上可能会多次阅读。同时该指标也属于后期加工指标 (Clifton Jr et al., 2007), 反映了人对于信息的重加工处理, 被试经过前期加工处理得到大致的主旨信息范围后, 再通过重加工以确定主旨小句。所以 *gist* 阶段的指标重要性排序也有较好的解释。

另外, 值得注意的是, 回视相关的指标是篇章理解过程的重要标志, 在 *article* 阶段, 第一遍阅读重要性相比第二、三遍阅读重要性更高, 在 *gist* 阶段, 最后一遍阅读重要性最高。最后一遍实际上指代的可能是第四或者第五甚至更多遍, 这反映了 *gist* 阶段对于兴趣区的回视次数多于 *article* 阶段。在心理学研究中, 阅读次数的增加有利于理解准确率的增加 (Schotter et al., 2014), 所以可以认为 *gist* 阶段被试对于句群的理解要好于 *article* 阶段, 这也从认知机理的角度提供了实验三中 *gist* 阶段的准确率要高于 *article* 阶段的原因。

最后, 对于重要性排序较为靠后的指标, *article* 和 *gist* 重合度较高, 例如第一个, 第二个, 第三个和最后一个注视点的注视时间等指标都较为靠后, 关于阅读兴趣区的注视点特性并不能和该句的重要性产生直接关联, 在阅读过程中, 上述几个指标的注视点都处于阅读的小句内部, 小句内部的注视点更可能反应的是小句内部的字或词语级别细粒度的信息重要程度, 例如小句内部不同字或词语对于该小句的重要性, 不能够反映小句对于整个语篇的重要性。实际上, 在心理学研究中, 以次为单位的注视时间是衡量以字或词为兴趣区的眼动指标, 而以遍为单位的注视时间则是衡量以短语或者句子为兴趣区的眼动指标 (朱滢, 2000)。对于不同细粒度而言, 研究者的研究重点是不同的, 像只有一个字或词的兴趣区, 研究者更关注于语言特征和词汇通达, 而对于包含短语或者小句的较大兴趣区而言, 研究者则关注词语的逻辑分析和句法的结构判断。重要性较低的指标启发我们对于句群的主旨信息提取而言, 人的认知过程是建立在 PCS 层小句为单位上的, 小句内部的眼动数据在此细粒度上不能够发挥作用。

图 4 给出了 *article* 和 *gist* 两个阶段的指标权重分布可视化情况。图中分为上下两部分, 分别为 16 列 5 行的像素矩阵, 5 行代表着随机选择的 PCS 层次片段, 16 列代表着 16 个指标, 矩阵中的颜色越深代表该特征越重要, 即特征重要性和矩阵块颜色深浅成正比。在图中可以看到, 对于 *article* 阶段而言, 整体的颜色分布大致为前六个左右较为深, 即总注视时间和次数以及第一遍相关的指标重要性较强, 第二、三注视时间的相关指标颜色最浅。对于 *gist* 阶段而言最后一

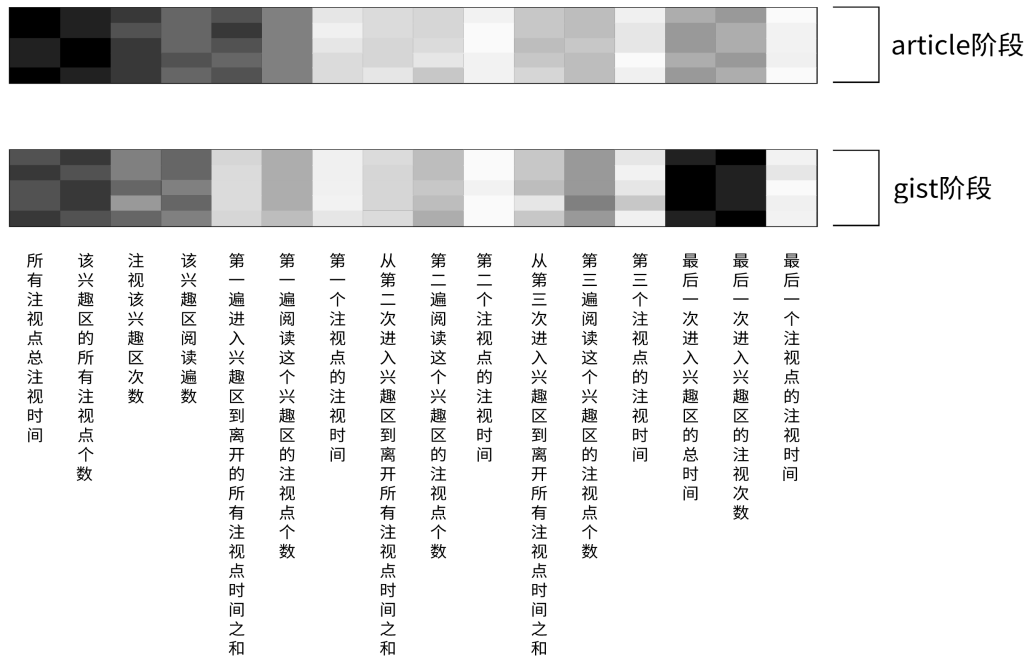


图 4: 不同阶段指标分布比较

次进入兴趣区以及总时间时间次数相关的指标颜色最深，重要性最强，同时颜色较深和颜色较浅的区域也和 *article* 阶段的区域有部分重合，也反映了两个阶段的相似性和相异性。该结果展示了两个阶段的指标不同显著性分布，凸显了自由阅读和主旨归纳阅读过程的眼动特征，也更加直观地感受到主旨归纳实际上更注重文本后期加工，以及两个阶段的重要指标从前期加工到后期加工的过渡。明确了主旨提取任务中的眼动模式。

5 结语

该文从眼动数据出发，结合人工主旨结构标注数据，验证了通过眼动数据建立主旨信息判断计算模型的可行性。

从线性到非线性的实验中可以看到眼动注视数据的复杂性，其更适合用非线性方法去处理和建模。在区分被试的实验中看到了不同被试的阅读水平有一定的差异性，对于主旨信息的理解也有一定的差别，高阅读水平的被试更有可能提取出更精确的信息。在最后的自由阅读和主旨归纳两个不同的阅读状态分析中，两个阶段的共同性和差异性反应出人在提取主旨信息时候的认知过程，并且此认知过程能够在眼动数据中找到支撑。

最终经过上述过程得到了0.76左右的PCS小句分类准确率，并且指标分析也有一定的认知解释，给出了主旨提取过程的眼动模式。据此可以知道对于文本主旨概括研究而言，结合眼动数据分析是一个可行的方法。未来从词语等不同的细粒度角度或者眼跳等不同的眼动模式角度分析处理眼动数据指标也是很重要的课题，同时对眼动数据的深度挖掘，结合文本语义模型构成多模态模型也是可以研究的方向。

致谢

论文工作得到国家自然科学基金重点项目61433018资助。中科院心理所博士生李琳主持完成了相关眼动数据收集和整理，中央民族大学4名本科生协助完成了相应短讯句群的主旨结构分析标注，在此一并表示感谢。

参考文献

- Hervé Abdi. 2007. Z-scores. *Encyclopedia of measurement and statistics*, 3:1055–1058.
- Irene Ablinger, Walter Huber, and Ralph Radach. 2014. Eye movement analyses indicate the underlying reading strategy in the recovery of lexical readers. *Aphasiology*, 28(6):640–657.
- Charles Clifton Jr, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. In *Eye movements*, pages 341–371. Elsevier.
- Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.
- John M Henderson and Fernanda Ferreira. 1990. Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(3):417.
- Jukka Hyönä and Robert F Lorch. 2004. Effects of topic headings on text processing: Evidence from adult readers' eye fixation patterns. *Learning and instruction*, 14(2):131–152.
- Jukka Hyönä, Robert F Lorch Jr, and Johanna K Kaakinen. 2002. Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology*, 94(1):44.
- Zhenchao Jiang, Lishuang Li, and Degen Huang. 2016. A general protein-protein interaction extraction architecture based on word representation and feature selection. *International Journal of Data Mining and Bioinformatics*, 14(3):276–291.
- Johanna K Kaakinen and Jukka Hyönä. 2010. Task effects on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(6):1561.
- Francis Quintal Lauzon. 2012. An introduction to deep learning. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 1438–1439. IEEE.
- Joseph Lee Rodgers and W Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.
- Quanzhong Liu, Chihau Chen, Yang Zhang, and Zhengguo Hu. 2011. Feature selection for support vector machines with rbf kernel. *Artificial Intelligence Review*, 36(2):99–115.
- Ya Lou, Yanping Liu, Johanna K Kaakinen, and Xingshan Li. 2017. Using support vector machines to identify literacy skills: Evidence from eye movements. *Behavior research methods*, 49(3):887–895.
- Olvi L Mangasarian and Gang Kou. 2007. Feature selection for nonlinear kernel support vector machines. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 231–236. IEEE.
- Christoph Molnar. 2020. *Interpretable Machine Learning*. Lulu. com.
- Mohamad T Musavi, Wahid Ahmed, Khue Hiang Chan, Kathleen B Faris, and Donald M Hummels. 1992. On the training of radial basis function classifiers. *Neural networks*, 5(4):595–603.
- Alexander Pollatsek, Keith Rayner, and David A Balota. 1986. Inferences about eye movement control from the perceptual span in reading. *Perception & Psychophysics*, 40(2):123–130.
- Keith Rayner, Jinmian Yang, Monica S Castelhana, and Simon P Liversedge. 2011. Eye movements of older and younger readers when reading disappearing text. *Psychology and aging*, 26(1):214.
- Keith Rayner. 1979. Eye guidance in reading: Fixation locations within words. *Perception*, 8(1):21–30.
- Erik D Reichle, Andrew E Reineberg, and Jonathan W Schooler. 2010. Eye movements during mindless reading. *Psychological science*, 21(9):1300–1310.
- SR Research. 2010. Eyelink 1000 user's manual, version 1.5. 2.
- Elizabeth R Schotter, Randy Tran, and Keith Rayner. 2014. Don't believe what you read (only once) comprehension is supported by regressions during reading. *Psychological science*, 25(6):1218–1226.

- Johan AK Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.
- Jiehang Xie, Xiaoming Wang, Xinyan Wang, Guangyao Pang, and Xueyang Qin. 2019. An eye-tracking attention based model for abstractive text headline. *Cognitive Systems Research*, 58:253–264.
- Yangtao Xue, Li Zhang, Bangjun Wang, Zhao Zhang, and Fanzhang Li. 2018. Nonlinear feature selection using gaussian kernel svm-rfe for fault diagnosis. *Applied Intelligence*, 48(10):3306–3331.
- 吴为章. 2000. 汉语句群. 商务印书馆.
- 周强. 2019. 微博短讯的主旨结构分析和质量检控研究. 清华大学信息技术研究院语音和语言技术中心技术报告, TH-RIIT-CSLT-TR20190809.
- 朱滢. 2000. 实验心理学. *MJ*. 北京大学出版社, 2004 年, 7.
- 李航. 2019. 统计学习方法第二版.
- 闫国利, 熊建萍, 臧传丽, 余莉莉, 崔磊, and 白学军. 2013. 阅读研究中的主要眼动指标评述. *心理科学进展*, 21(4):589–605.

JCL2020