Journal of Computer Applications
计算机应用,2020,40(8):2171－2181
ISSN 1001-9081
CODEN JYIIDU
2020-08-10
http://www.joca.cn

文章编号:1001-9081(2020)08-2171-11
DOI:10.11772/j.issn.1001-9081.2019122204

基于注意力网络的语体多元特征挖掘

吴海燕,刘　颖*
(清华大学 人文学院,北京 100084)
(*通信作者邮箱yingliu@mail.tsinghua.edu.cn)

摘　要:针对大规模语料中不同语体的特征难以挖掘、需要大量专业知识和人力的问题,提出了一种自动挖掘能区分不同语体的特征的方法。首先,将语体表示成词、词类、标点符号、它们的2元、句法结构及多种组合特征;然后,使用注意力机制和多层感知机(MLP)的组合模型(如注意力网络)把语体分类成小说、新闻和课本,并在过程中自动地提取出能够帮助区分语体的重要特征;最后,通过对这些特征的进一步分析,可以得到不同语体的特点及一些语言学结论。实验结果显示,小说、新闻和课本在词、主题词、词的依存关系、词类、标点符号和句法结构都有显著的差异,进一步表明了人们在使用语言时因交际对象、目的、内容和环境的不同,对词汇、词类、标点和句法的运用上会自然地呈现出某种不同。

关键词:语体特征挖掘;语体特征区分度;注意力机制;多层感知机
中图分类号:TP391.1　　**文献标志码**:A

Stylistic multiple features mining based on attention network

WU Haiyan, LIU Ying*
(*School of Humanities*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: To solve the problem that it is difficult to mine the features of different registers in large-scale corpus and it needs a lot of professional knowledge and manpower, a method to mine the features of distinguishing different registers automatically was proposed. First, the register was expressed as words, parts-of-speech, punctuations, and their bigrams, syntactic structure as well as multiple combined features. Then, the combination model of attention mechanism and Multi-Layer Perceptron (MLP) (i.e. attention network) was used to classify the registers into novel, news and textbook. And, the important features that were able to help to distinguish the registers were automatically extracted in this process. Finally, through the further analysis of these features, the characteristics of different registers and some linguistic conclusions were obtained. Experimental results show that novel, news, and textbook have significant differences in words, topic words, word dependencies, parts-of-speech, punctuations and syntactic structures, which implies that there will naturally present some diversity in the use of words, parts-of-speech, punctuations, and syntactic structures due to the different communication objects, purposes, contents, and environments when people utilize language.

Key words: stylistic feature mining; discrimination measure of stylistic feature; attention mechanism; Multi-Layer Perception (MLP)

0　引言

语体是当代语言学的重要范畴,是语言研究的一个不可忽视的领域,它是人们在使用语言时受到交际对象、目的、内容、环境等交际条件的限制而形成的一些综合的语言特点。霍小立[1]指出,语体特征是语体在受交际环境、目的和内容影响而间接体现语体本质的属性集合。顾名思义,语体特征区分度是指特征能区分不同语体的能力。

语体特征作为语体的本质属性,具有一定的语体区分度。在之前的文献中主要从两方面进行语体区分度的研究:一是根据语言学知识分析哪些特征具有语体区分度;二是借助计算机模型,通过分析分类准确率或聚类离散程度等来说明哪些特征具有显著语体区分度。根据以往研究方法的不同,本文将它们分为三类:语言学方法、统计学方法及神经网络方法。

1)语言学方法。很多语言学者对不同语体中具体的字、词、特定短语、句法结构、句类等都做了详细的分析研究,并给出了哪些特征具有语体区分度。例如:2010年陶红印等[2]提出"把"字句和"被"字句在不同的语体中具有明显的差异。冯胜利[3]通过对比研究口语和书面语的语体特征,最后指出单、双音节是区别口语和书面语的基本单位。张豫峰[4]指出"得"在文艺语体、政论语体、科技语体和公文语体中频率呈递减趋势。钱小飞[5]认为"地"字结构可用来区分不同的语体。句法作为汉语的语言结构组成之一,方梅[6]认为句法特征在不同语体的分布存在差异,它在宏观上规定的句子语气类型和功能类型也存在差异。此外,标点符号也具有语体区分性,林毓

收稿日期:2020-01-02;**修回日期**:2020-03-16;**录用日期**:2020-03-18。
基金项目:国家社会科学基金资助项目(18ZDA238);教育部人文社科一般项目(17YJAZH056);北京社会科学基金资助项目(16YYB021)。
作者简介:吴海燕(1985—),女,陕西延安人,博士研究生,CCF会员,主要研究方向:自然语言处理;　刘颖(1969—),女,内蒙古赤峰人,教授,博士,CCF会员,主要研究方向:语料库语言学、计算语言学、机器翻译。

霞[7]曾指出标点符号的运用同语体有着密切的关系，标点符号种类的多寡、频率的高低取决于语体的形式。

这些研究的共同点是通过语言学分析和基本的计量找出哪些字、词、短语或句法结构等特征具有语体区分度，并对其重要性进行解释说明。这些语言学家细致入微的语言学分析为后续学者的研究提供了重要的理论依据，但这些分析说明是从语体内在的含义出发，需要逐词逐句地分析和筛选例句，工作量大，耗时费力，并且缺乏大规模数据上的统计验证。

2）统计学方法。随着计算机技术的发展，学者们逐渐开始借助统计学方法来提取语体特征并将其数字化表示，例如：频率、词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）等形式，并在此基础上进行分类和聚类，通过分析分类和聚类的准确率来说明所选语体特征的重要性。胡骏飞等[8]借助词频来分析“弄”字句在会话口语、影视口语及书面语体中的分布，并得出词类在一定程度上可以反映语体的重要程度，同时从语用功能的角度加以解释说明。肖天久等[9]利用词和词的N元文法相结合研究《红楼梦》前八十回与后四十回的关系，结论是前八十回与后四十回有差异，这说明词和词的N元文法对判定语体有重要的作用，即词和N元文法具有语体区分度；但是该文作者并未直接给出这个区分语体能力的大小，而是通过使用聚类的离散程度来间接说明的。

这些方法的共同点是对所提取的特征都采用了数值量化表示，与语言学方法相比，特征的形式化表示很容易被计算机处理，并且也取得了很好的效果。事实上，从这些研究所选择的特征类型来看，学者们已经开始尝试从更多的角度考虑特征的选择，唯一不足的是这些特征需要人为选取，这就对研究人员的语言学知识有较高的要求。

3）神经网络方法。近些年来，基于神经网络的学习方法在很多领域都有出色的表现。利用神经网络来挖掘语体的重要特征也是学者们努力研究的方向。周浩[10]利用神经网络对句法结构进行分析，实验结果显示与传统方法相比，神经网络能挖掘更好的句法结构来帮助分析语体。还有学者利用复杂神经网络挖掘语体的关键信息，例如Wang等[11]提出了一种基于双向长短时记忆（Long Short-Term Memory，LSTM）和循环神经网络（Recurrent Neural Network，RNN）的关键词自动提取方法。通过对京东的产品评论进行分析，所提取的关键词能很好地区分不同种类的产品。此外，神经网络在其他领域也有很好的应用，Bahdanau等[12]将注意力机制应用到翻译领域，通过注意力机制找出每一句话的关键词来帮助终端进行翻译，从而使翻译效果得到了很大的提升。Pappas等[13]将注意力机制应用到词和句子层面上，通过注意力机制找到能区分文本的关键词和句子，从而提高了文档分类的准确率。

这些基于复杂神经网络的方法无论是在特征提取还是分类准确率方面都有了很大的提升。深度学习可以自动获取语体本身的信息，并使用高维度的向量来表示所提取的特征，使其具有高维度的空间语义属性[14]。这样，语体特征的语义信息被挖掘出来，同时也极大地减少了人为参与。不足的是这些模型对硬件要求比较高，训练时间通常比较长。

通过分析这些方法的优缺点，本文利用注意力机制和多层感知机的组合模型——注意力网络来挖掘能区分不同语体的重要特征，结构如图1所示。本文工作的核心是通过注意力网络挖掘能区分小说、新闻及课本的词、词类、标点符号、句法结构及它们的2元特征，并对它们的重要性进行量化表示与分析，主要工作是：

1）利用注意力网络模型能挖掘出多种能区分小说、新闻及课本的特征集，主要包含：词和词的2元、词类和词类2元、标点和标点2元、句法结构以及多种特征组合，从多个语言层面上挖掘出小说、新闻和课本在词汇、句法和语义使用上系统的差异。

2）从语义角度挖掘出小说、新闻及课本的主题词及与其依存的从属词和句法结构；并进一步挖掘出动词是三种特征内在联系的纽带，同时证明了动词对语体的重要性；最后，还逐一挖掘了小说的主人公形象（身体器官、面部表情、内心活动、说话语气、亲属称呼、社会角色等）、新闻的事件报道（时间、地点、主题等）及课本的人物描写和议论主题等的内在联系。

3）对注意力网络进行改进，使得它不但能够挖掘序列化的特征，也能挖掘非序列化的句法结构；而且它能很好地挖掘出区分小说、新闻及课本的句法结构集。

4）本文选用注意力网络模型的优点是：能够挖掘出多种有效特征，同时还能给出每一种特征的语体区分度；能够自动过滤掉大量的冗余特征；能够自动过滤掉停用词。

1 研究方法

1.1 注意力网络

在计算机领域，深度学习已成为一种流行的方法，它在多个领域都显示出了强大的建模能力[15]。在神经网络的模式设计中，有一种结构被称为注意力机制，它能自动分析文本中不同信息的重要性。在自然语言处理领域，学者们将它与深度模型相结合已经取得了显著的成果。

本文借助注意力机制来挖掘具有显著区分度的语体特征。这些特征主要包括：词的N元、词类的N元、标点符号的N元及句法结构。首先，通过对由这些特征所表示的文本执行分类训练，在训练的过程中，注意力机制会对这些特征进行评分。这里，注意力机制的作用是找出哪些特征具有显著的语体区分度并赋予相应的注意力分值，分值越高就越能区分语体，即注意力机制分值越高该特征的语体区分度就越大。本文使用的注意力网络结构如图1所示，主要由输入层、嵌入层、N元向量层、注意力层、N元句子向量层、连接层、分类层及输出层组成。

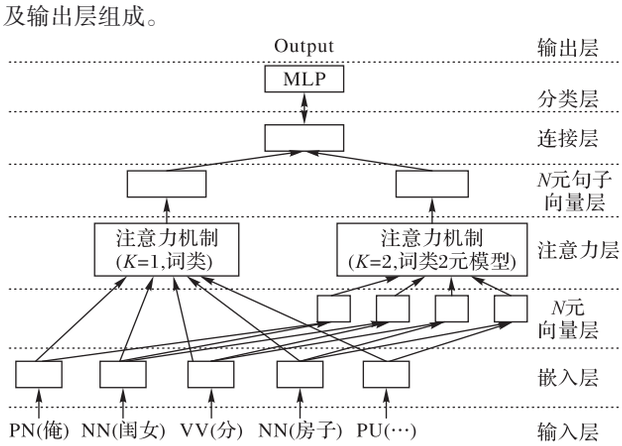


图1 注意力网络结构

Fig. 1 Structure of attention network

图1示意的是将词类或词类的2元作为输入时的网络结构。在两种情况下，均先将输入的词类通过嵌入层转化为词类向量。若对词类1元（即词类）进行评分，则词类向量直接输入到注意力层；若对词类的2元进行评分，则词类向量先通

过N元向量层组合产生N元向量,再输入到注意力层。最终,无论特征是词类还是词类的N元,注意力网络层将对输入到全连接层的句子进行分类。总的来说,图1注意力网络结构包含三个部分:模型特征输入(输入层、嵌入层及N元向量层)、注意力机制特征评分(注意力层和N元句子向量层)及语体分类(连接层和分类层)。接下来,将详细地介绍这三部分。

1.2　模型特征输入

本节的主要目的是将由句子组成的语料集转换成注意力网络所要识别的特征(词、词类、标点符号、句法结构及它们的组合)向量,主要包括输入层、嵌入层及N元向量层。

1)输入层。使用模型前,需要用特征表示语料集中的每一个句子。首先,需要构建特征对应的字典$\boldsymbol{W}=\{w_1,w_2,\cdots,w_n\}$,$n$表示文本的特征数。例如:想提取能区分小说、新闻及课本的词汇特征,故此时的$\boldsymbol{W}$是所有词的集合,n表示不同词的数目。类似地,如果想要提取能识别小说、新闻及课本的词类或标点符号或句法结构特征,此时的$\boldsymbol{W}$就是词类或标点符号或句法结构,对应的n就是这几类特征各自的总个数。其次,将语料集的句子集用字典$\boldsymbol{W}$中的特征表示,其中,L表示语料集的句子数,即$S=\{s_1,s_2,\cdots,s_L\}$。对每一个句子进行切词、词性标注及构建句法树如下:

$$S_{i\text{-words}}=\{w_{i,1},w_{i,2},\cdots,w_{i,n}\}$$
$$S_{i\text{-POS}}=\{pos_{i,1},pos_{i,2},\cdots,pos_{i,m}\}$$
$$S_{i\text{-syntax}}=\{sy_{i,1},sy_{i,2},\cdots,sy_{i,p}\}$$

其中:$pos_{i,j}$表示词$word_{i,j}$所对应的词性,i表示句子在语料库中的序号,j表示该词类在当前句子中的序号;m为句长;p是句法树经过序列化处理(前序遍历,即先访问根节点,然后访问左子树,最后访问右子树)后所得的句法结构数。对于句法结构的提取,需要借助句法树来完成。

下面利用图1的例句来详细说明以上几种形式化表示。首先,使用斯坦福自然语言处理工具包CoreNLP对句子$s_i=$\{俺闺女分房子…\}分别进行切词、词性标注及构建句法树(图2)得:

$$S_{i\text{-words}}=\{\text{俺 闺女 分 房子 …}\}$$
$$S_{i\text{-POS}}=\{\text{NR, NN, VV, NN, PU}\}$$
$$S_{i\text{-Pun}}=\{\cdots\}$$
$$S_{i\text{-syntax}}=\{\text{IP}\rightarrow\text{NP VP PU, NP}\rightarrow\text{PN NN, VP}\rightarrow\text{VV NP,}\ \text{NP}\rightarrow\text{NN}\}$$

语料集中所有的句子分别用类似$s_{i\text{-words}}$、$s_{i\text{-POS}}$、$s_{i\text{-Pun}}$及$s_{i\text{-syntax}}$表示后输入到嵌入层。

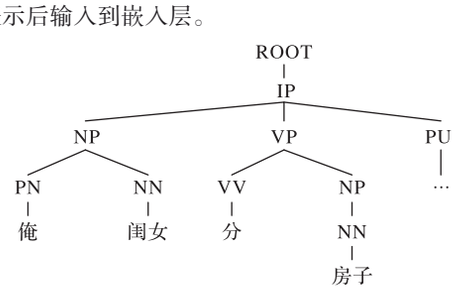


图2　句法树

Fig. 2　Syntactic tree

2)嵌入层。该层是将句子特征转化为向量,以词特征为例,即:

$$\varphi(w_{i,j})=\boldsymbol{w}_{i,j}$$

其中:$\boldsymbol{w}_{i,j}\in\mathbf{R}^v$为词$w_{i,j}$所对应的向量(本文用粗体表示相应特征的向量);$\varphi$为特征空间到向量空间的映射,即$\varphi:\boldsymbol{W}\rightarrow\mathbf{R}^v$,$v$表示特征向量的维度,由3.2节实验设置给出。词向量由正态分布N(0, 0.01)随机初始化得到,在神经网络训练过程中会被训练优化。

3)N元向量层。该层是将嵌入层所得的特征向量按照N的大小拼接起来,以词类的N元为例:对于句子$S_{i\text{-words}}=\{w_{i,1},w_{i,2},\cdots,w_{i,n}\}$所对应的词类表示为$S_{i\text{-POS}}=\{pos_{i,1},pos_{i,2},\cdots,pos_{i,m}\}$,则词类的$k$元表示为$S_{i\text{-tag}}=\{g^k{}_{i,1},g^k{}_{i,2},\cdots,g^k{}_{i,m-k+1}\}$,其中,$g^k{}_{i,j}$表示句子的第$j$个词类$k$元,用粗体表示其向量,则它对应的向量是:

$$\boldsymbol{g}^k{}_{i,j}=\left\{\boldsymbol{pos}^k{}_{i,j}\middle|\boldsymbol{pos}^k{}_{i,j+1}\middle|\cdots\middle|\boldsymbol{pos}^k{}_{i,j+k-1}\right\}$$

其中:“|”表示连接,向量$\boldsymbol{g}^k{}_{i,j}$就是词类的N元向量。对于由词、标点符号及句法结构经过嵌入层后所得的向量作相同的操作,就可以得到这几类特征对应的N元向量。

经过模型输入部分得到句子特征的N元向量,接下来需要利用注意力机制对其进行评分。以图1词类表示的句子为例,来阐述注意力机制的评分原理。

1.3　注意力机制对特征评分

1)注意力层。首先,注意力机制通过全连接层计算出每一个句子的第j个词类k元特征($g^k{}_{i,j}$)的注意力向量。

$$\boldsymbol{u}^k{}_{i,j}=\tanh(\boldsymbol{A}^k\boldsymbol{g}^k{}_{i,j}+\boldsymbol{b}^k)\tag{1}$$

其中:$\boldsymbol{A}^k\in\mathbf{R}^{t\times kv}$和$\boldsymbol{b}^k\in\mathbf{R}^t$是注意力网络的参数,分别为连接权重和偏置,$t$表示注意力网络的隐含层的维度,$v$表示向量的维度,$kv$表示向量$\boldsymbol{g}^k{}_{i,j}$的维度。其次,因为Kalman等[15]曾经指出“具有非线性多项式激活函数的多层前馈网络可以逼近任何函数”,因此为了使模型具有更好的拟合性,通常在全连接层之后增加一个非线性多项式激活函数。其中,$\boldsymbol{u}^k{}_{i,j}$是包含词类k元模型(k-Gram)重要性信息的隐藏注意向量。之后对注意力隐含向量进行加权求和,公式如下:

$$\boldsymbol{u}^k{}_{i,j}=\boldsymbol{h}^{k\mathrm{T}}\boldsymbol{u}^k{}_{i,j}\tag{2}$$

其中:$\boldsymbol{h}^k$是权重,属于注意力网络参数;$\boldsymbol{u}^k{}_{i,j}$是注意力机制给k元的$\boldsymbol{g}^k{}_{i,j}$所打的分值。注意,$\boldsymbol{u}^k{}_{i,j}\in(-\infty,\infty)$,如果直接用$\boldsymbol{g}^k{}_{i,j}$的分值与其对应的特征向量进行加权求和来形成句子向量,那么随着训练过程的进行,句子向量的长度和规模将失去控制趋向无穷大。所以,需要对句子向量进行归一化,本文使用函数规范化指数函数Softmax函数进行归一化。该函数将$m-k+1$个实数作为输入,并将其规范化为概率分布,公式如下:

$$a^k{}_{i,j}=\exp(\boldsymbol{u}^k{}_{i,j})\Big/\sum_j\exp(\boldsymbol{u}^k{}_{i,j})\tag{3}$$

即$a^k{}_{i,j}\in(0,1)$,且$\sum_j a^k{}_{i,j}=1$。$a^k{}_{i,j}$表示第i个句子的第j个位置的词类k元模型的注意力权重。所有句子的词类N元模型的注意力分值都可以通过式(1)、(2)和(3)来表示。如图1中右侧部分为词类2元的注意力机制。换句话说,式(1)~(3)对应图1的注意力层。

2)N元句子向量。通过式(1)~(3)完成了对句子中词类的N元的评分。这样就可以将原来的句子向量表示为带有注意力分值的词类和词类N元模型,如式(4):

$$s^k{}_i=\sum_j a^k{}_{i,j}\boldsymbol{g}^k{}_{i,j}\tag{4}$$

一般来说,这里的句子向量是注意力分值所有向量以权重加权和所得,它的权重是随着注意力网络的训练动态生成的,不同句子的词类N元模型的权重是不一样的。注意力网络会随着训练分类准确率的提升动态地为每一个词类N元模型进行评分。经过注意力层和N元句子向量化表示后,得到了带有注意力分值的句子向量。接下来需要使用分类器对这

些句子进行分类。

1.4 语体分类

本文使用多层感知机(Multi-Layer Perceptron, MLP)对语体进行分类。MLP是一种前馈人工神经网络,一般由输入层、隐藏层和输出层组成,每层都有很多个神经元。MLP通过使用后向传播的有监督算法来训练和学习区分不同的语体。本文以句子向量$\boldsymbol{s}_i$为输入,返回不同语体的概率作为输出。假设C是所有不同语体的集合,$|C|$是语体的数目。

1)连接层。本文通过使用两个完全连接的层来构建一个高效简单的分类模块,公式如下:

$$\boldsymbol{p}_i = \boldsymbol{M}_2 \tanh(\boldsymbol{M}_1 \boldsymbol{s}_i + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{5}$$

其中:$\boldsymbol{M}_1 \in \mathbf{R}^{t \times v_s}$,$\boldsymbol{b}_1 \in \mathbf{R}^t$,$\boldsymbol{M}_2 \in \mathbf{R}^{|C| \times t}$,$\boldsymbol{b}_2 \in \mathbf{R}^{|C|}$,这四个参数都是模型参数,$v_s$是句子向量$\boldsymbol{s}_i$的大小,$t$是隐含层的大小,$|C|$是语体类别个数。$\boldsymbol{p}_i$向量表示句子属于不同语体的非规范化概率,其中$p_i^{(j)}$为向量的第$j$个数表示句子$s_i$属于语体$j$的非规范化概率,本文使用如下函数进行归一化:

$$p(c_j|s_i) = \exp(p_i^{(j)}) \Big/ \sum_j p_i^{(j)} \tag{6}$$

其中,$p(c_j|s_i)$表示句子s_i属于类别c_j的概率。在本文类别指的是小说(0)、新闻(1)及课本(2)这三类。

2)分类层。为了给出预测类别,选取最大$p(c_j|s_i)$所对应的类别c_j作为模型预测类别,这就是图1中的分类层。

以上涉及的训练参数会在3.2节的实验设置中逐一给出。

另外,对于组合特征("词+词类"、"词+标点符号"、"词+词类+标点+句法结构")来说,由于词类(32种,具体含义见表12)、标点符号(12种)及句法结构(高频的396种)的数量比较少,采用One-Hot编码表示,并取它们与词嵌入向量的和表示组合特征向量。对于这几类组合特征向量,只需用图1的左边的模型重复上面的步骤即可。

本文使用最常见的12种标点符号,即,句号(。)、感叹号(!)、问号(?)、省略号(……)、逗号(,)、顿号(、)、分号(;)、引号(“ ” ‘ ’)、冒号(:)、括号(( ) [ ] { })、破折号(——)和书名号(《》〈〉)。

2 研究过程

本文的研究过程由以下几个步骤组成:

1)构建语料库。本文的研究对象是小说、新闻及课本,具体信息在3.1节语料库介绍中详细说明。

2)语料预处理。本文语料的处理使用斯坦福大学所提供的自然语言处理工具包Stanford CoreNLP进行,主要包括数据清洗、切词、词性标注、句法树构建等。其中,语料库的处理以句子为单位,判断句子的标准是以号(。)、问号(?)、感叹号(!)及省略号(……)为结尾的句子。

3)给每一个句子编号。通过建立特征字典,将语料库中每一个句子所对应的特征用其在字典中唯一的编号来表示,进而将语料库中所有的句子转换为用特征编号来表示。

4)注意力机制和多层感知机组合模型。这是注意力网络的核心部分,其中,注意力机制对输入句子进行评分,其分值的大小随着分类准确率的变化而自动调整,直到分类准确率达到最优而停止更新。而多层感知机是一个分类器,用于对句子类别的预测。

5)单现、共现处理。无需计算特征出现在每一种语体的次数,对每一种语体中的所有特征求注意力分值的平均分值。

6)特征选择。通过绘制注意力网络分值的分布曲线,找出每一种特征所对应的注意力分值的阈值,进而选择出能区分小说、新闻及课本的关键特征。

3 实验结果与分析

3.1 语料库

本文选取小说、新闻及课本三种语料,具体信息如下:

1)小说。选取莫言和余华的小说,其中包括莫言的12部小说:《白棉花》《丰乳肥臀》《红高粱》《红树林》《酒神》《生死疲劳》《十三步》《食草家族》《四十一炮》《檀香刑》《天堂蒜薹之歌》及《蛙》;余华的8部小说:《第七天》《古典爱情》《活着》《现实一种》《兄弟》《兄弟2》《许三观卖血记》及《在细雨中呼喊》。

2)新闻。选取搜狗公开的语料集(https://www.sogou.com/labs/resource/list_yuliao.php),主要包含国内外新闻、财经、股票、房地产、健康、热点、教育及社会等十个主题相关的新闻。

3)课本。以中小学的语文教材为主,包括国内外小说、散文、励志故事、爱国故事、话剧等,例如:鲁迅的《孔乙己》《阿Q正传》《祥林嫂》及《故乡》等;海明威的《海燕》;莎士比亚的《罗密欧与朱丽叶》;朱自清的散文《背影》及《匆匆》等。由此可以看出,课本包含的语体种类比较多,其目的通常是选取一些有代表性的文章来培养学生的听说读写等能力。数据集详细的统计信息见表1。

表1 数据集信息

Tab. 1 Dataset information

语料	总句子数	训练集句子数	验证集句子数	测试集句子数
小说	117 353	93 882	11 736	11 735
新闻	29 754	23 803	2 796	2 795
课本	24 119	19 295	2 412	2 412

3.2 实验设置

实验将语料集按8:1:1划分为训练集、验证集和测试集,验证集用来探索训练轮数且在过拟合的情况下提前结束训练。为了更好地训练模型,本文使用网格搜索来选择模型参数的最优组合,这些参数主要包括:学习率(learning rate)$\in$ {0.001, 0.01, 0.1, 1}和批量大小(batch size)$\in${32, 64, 128, 256, 512},初始化向量的维度是128。另外,本文实验以句子为单位进行训练分类,故需要设置句子长度及每个词用多少位来表示。小说和课本的平均句子长度接近20,新闻的平均句子长度接近30。因此,设置句子长度集$\in${10, 20, 30, 40, 50, 80, 100, 120, 130}。句子向量的大小是这三种语体特征的总数,特征的维度大小设置为32。参数的最佳组合以黑色加粗显示,对模型影响较小的其他参数则统一采用默认值。对于用句法结构表示的句子,在训练时将句子长度大小改为200,其他参数不变。本文采用准确率来评估模型的性能。

3.3 结果及分析

通过回答以下2个问题进行实验结果分析。

1)问题1:对词、词类、标点符号及句法结构来说,当注意力分值为多大时才能很好地区分小说、新闻及课本。

以训练词特征的结果分析为例,将其注意力分值按照降序排列,然后取队尾、队首词进行分类,其准确率随取队首、队尾的词的多少而变化,其变化曲线(包含训练集)如图3所示。

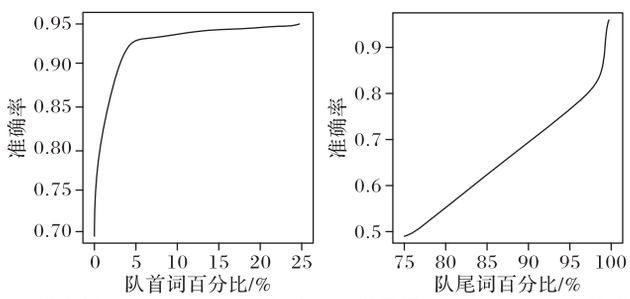
根据图3分析如下:

1)从图(a)的队首词比,大约用队首3%的高注意力分值词就能使模型的分类准确率达到90%以上,表明高分值的词具有非常好的语体区分度。

2)从图(b)的队尾词可以看出,大约用队尾97%的低注意力分值的词才能使模型分类准确率达到90%以上,表明低分值的词对区分语体的帮助没有高注意力分值的词好。

3)在一定程度上,无论是取队尾词还是队首词,有效的特征越多,其分类准确率越高。

上述结果验证了注意力分值具有很好的区分度,根据不同注意力分值词的百分比和其对应的准确率,本文将注意力的分值分为高([0.15, 1])、中([0.01, 0.15))、低([0, 0.01))三个区间,不同区间的词频占比及其对应的准确率见表2所示。从表2可以看出,低区分度的词占大多数(约75%)所对应的分类准确率只有47.60%;而取高分值词的4.21%,对应的分类准确率就达到93.31%。这说明在区分不同的语体时,高分值的词更有效。同时也说明了研究语体特征的意义:挖掘更多的具有高注意力分值的特征来提高语体分类准确率,进而实现语体特征的降维。



(a) 队首词百分比与准确率的关系　(b) 队尾词百分比与准确率的关系

图3　队尾与队首词百分比与准确率的关系

Fig. 3　Relationship between accuracy and proportion of head/tail words of queue

表2　词的分值区间、频率及准确率的分布

Tab. 2　Distribution of score interval, frequency and accuracy of words

分值区间	频率占比/%	准确率/%
[0.0,0.01)	74.99	47.60
[0.01,0.15)	20.80	78.54
[0.15,1.0]	4.21	93.31

2)问题2:对词、标点符号、词类及句法结构来说,每一种特征区分小说、新闻及课本的能力如何。

分别使用词、词类、标点符号、句法结构及它们的组合特征表示语料,并将其作为输入特征,经过训练后得到的分类结果如表3所示。

表3　基于语体特征的分类结果　　单位:%

Tab. 3　Classification results based on stylistic features　unit: %

语体特征	准确率	语体特征	准确率
词	91.22	词类2元	78.96
句法结构	80.94	词+词类	91.25
词类	78.33	词+标点符号	91.39
标点符号	77.54	词+词类+标点符号	92.54
词2元	93.04	词+词类+标点符号+句法结构	93.33
标点2元	78.82		

根据表3的分类结果可以得出以下几点:

1)对于每一种特征(词的N元、词类的N元、标点符号的N元及句法结构)来说,分类的准确率由高到低依次是:词的2元、词、句法结构、词类的2元、标点符号的2元、标点符号及词类,这几类特征都具有语体区分能力,但是每一种特征能区分小说、新闻及课本能力的大小并不相同。总体来说,词和词的2元的分类准确率相对比较高,这是因为相比较词类、标点符号及句法结构,词是最小的能够独立活动的有意义的语言成分,且具有实际含义。词的2元特征是词的组合,所以比词含有更丰富的信息,因此词的2元分类准确率最优。句法结构表示词之间搭配规则,是词语组成句子的必要结构,由它构成的词组既可以单独成句,也可以是句子的组成成分。所以从这个角度来说,句法结构具有较高的语体区分度。标点符号不但具有表示句子停顿、结束等功能,还可以表达句子的语气,尤其是句末标点符号(感叹号、疑问号、省略号)等。然而,对于小说、新闻及课本来说,句子的语气特征十分重要,而词类的作用仅是指明词的性质,所以与标点符号相比,词类语体区分度没有标点符号的好。但是,从表3的分类结果来看,标点符号的2元没有词类的2元的分类效果好,一方面是因为词类的种类(32)比标点符号的种类(12)多,所以词2元的组合特征比标点符号2元的组合特征多,这就会导致基于标点符号2元训练的注意力网络处于欠拟合,没有达到最优状态,故其效果不好;另一方面,词类的2元从某一种角度上来说,体现了词之间的搭配共现规则,尤其是那些高频率的词类的2元。同样,根据表3,作为表示词之间搭配规则的句法结构来说,基于它的分类准确率高于标点符号,这说明词之间的搭配规则在区分语体上也有重要的作用。所以结合这几点,词类的2元比标点符号的2元更具有语体区分度是合理的。

2)对于组合特征来说,基于"词+词类+标点符号+句法结构"的分类效果最优,其次是"词+词类+标点符号",最后是"词+词类"。反过来看,每增加一类特征,所对应的分类准确率就有所提高,只是提高的程度有所不同,所以说每一类特征都具有语体区分度。这是因为每一类特征都是从不同的角度分析语体。这样通过多类组合特征,就可以从多个角度区分语体,并根据其对应的准确率能很好掌握每一类特征对区分语体的影响。更进一步说明了综合考虑多种特征能够更有效地区分开不同语体。

接下来,用一个例子分析注意力分值在不同语体特征上的分布情况。选用基于"词+词类+标点符号"训练后所得的注意力分值分布如图4所示。在图4中,分别选取了长度差不多的4个句子,其中,第一句来自小说(余华的《古典爱情》),第二句选自新闻(《上海滑稽剧团的近况》),第三句和第四句选自课本(《修辞手法》和秦似的《榕树的风度》),之所以从课本中选取两句是因为课本所包含的语体种类比较多,这样可以进一步了解注意力分值在不同语体总的分布情况。

图4中灰度越深表示该特征的注意力分值越高,即该特征越重要。第一个句子的"柳生"颜色最深,根据右边的注意力分值刻度值,发现其注意力分值大于0.15,所以"柳生"是这句话的关键词,且符合该句的语义描述。我们知道,"柳生"是余华的小说《古典爱情》的主人公,该文全篇都是以"柳生"为主展开叙述的。同理,第二句来自新闻,是一篇有关于《上海滑稽剧团的近况》的报道,讲述了"滑稽剧团"从产生、发展、兴盛到衰败的过程,从而感慨任何事物都要经历这样的过程。故其关键词是"滑稽"和"剧团"。第三句是关于修辞方法的议论分析,故其关键词是"修辞"。第四句,根据上下文含义,该句是作者看见榕树在艰苦的环境中依然茁壮成长有感而发,并通过一个疑问句来强调"这个时候"榕树十分美丽。由此可见,注意力网络很好地学习到了这种情况下作者想表达的含义并对其进行准确的评分。

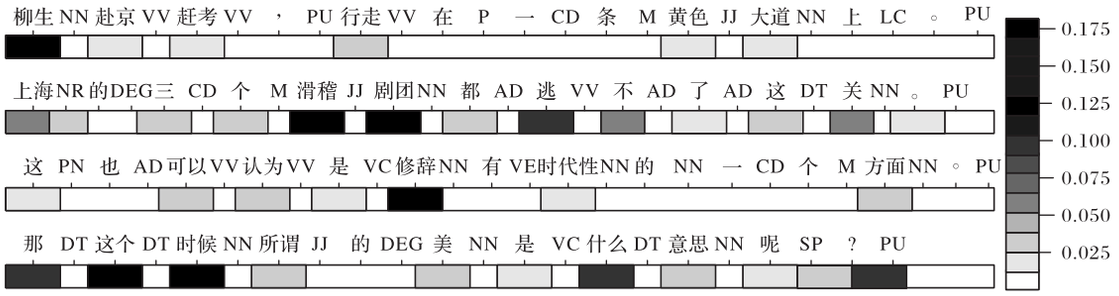


图4 注意力分值分布

Fig. 4 Distribution of attention score

4 特征语体区分度分析

4.1 词的语体区分度统计和分析

词是最小的语言运用单位，且能独立表达完整的意思。根据3.3节的问题1，选择满足条件的前几个高注意力分值的词进行分析，高分值的词如表4所示。从表4可以看出，小说的高分值词大部分都是小说主人公的名字；新闻的词主要是主题词，如热点、房价、股市，还有一些较为正式的词，如表决、议案等；课本的关键词是小说选篇的主人公的名词、人物传记名词等。为了进一步分析小说、新闻及课本词的差异，下面将从词的语义信息和词之间的依存关系进行深入分析。

表4 高注意力分值的词

Tab. 4 Words with high attention score

语体	高注意力分值的词
小说	蒜薹、上官、李光头、老兰、柳生、许三观、整容、小魏、姑姑、高密、金菊、司马库
新闻	热点、议案、传销、房价、表决、股市、获利、公告、消费者、指标、资费、目前、全球
课本	阿Q、齐仰之、陈毅、翠翠、陈奂生、列宁、太太、孔子、喜儿、死海、闰土、罗密欧

4.1.1 主题词分析

为了进一步分析小说、新闻及课本的关键词，使用T分布随机近邻嵌入(t-distributed stochastic neighbor embedding，t-SNE)降维算法将所提取关键词的向量映射到二维平面内表示，并选择每一个簇中注意力分值最高的词作为该簇的语义主题词，如表5所示。表5中，小说的主题词主要是主人公的名字(柳生、余占鳌)及地点名词(高密、东北)为主；新闻主要是事件主题名(股市、经济、市场等)及核心人物(主席)等；课本的主题词是人名(高尔基、列宁)、小说选篇的主人公名字(孔乙己、闰土)、议论文的主题词(爱国)等。

表5 语义主题词分布

Tab. 5 Distribution of semantic topic words

语体	语义主题词
小说	柳生、余占鳌、高密、福贵、姑姑、东北……
新闻	股市、议案、经济、主席、市场、文化……
课本	孔乙己、列宁、闰土、孔子、高尔基、爱国……

4.1.2 主题词的支配词分析

依存关系表示句子中两个词之间的2元关系，其中一个为核心词，另一个为依存词，反映的是核心词和依存词之间语义上的依赖关系。在不同的语体中，词与词之间的依存关系是否存在差异？已有研究[16]证明了依存句法关系能很好地区别不同的作者。本文挖掘主题词与其支配词之间的依存关系并按降序排列，结果如表6所示。

表6 小说主题词的支配词分布

Tab. 6 Distribution of governing words of the topic words in novel

关系名	个数	支配词
nsubj	63 526	赶赴、呻吟、抱、询问、颤抖
amod	21 791	异常、静静、开心、无助、迷茫
dobj	21 112	说、笑、跑、跳、想、跑、喊、叫
poss	14 506	眼睛、鼻子、嘴、脸、爹、娘、弟弟、胸膛、回想、精神、感觉

以小说的主题词“柳生”为例，选择包含主题词“柳生”的句子：“柳生 赴京 赶考 ，行走 在 一 条 黄色 大道 上 。”建立相应的依存树，如图5所示。

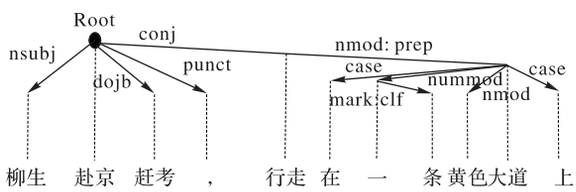


图5 依存树

Fig. 5 Dependency tree

同1.2节一样，该句的依存树也是调用斯坦福自然语言处理包完成的。对于图5中词之间的相互依存关系用如下形式表示：

依存关系名(支配词$_{位置}$，从属词$_{位置}$)

这里的“依存关系名”由斯坦福自然语言处理包中的依存句法关系给出，一共53个。“从属词$_{位置}$”是在依存句法树中箭头的结束词(从属词)，“位置”表示该词在句子中的位置；相反“支配词$_{位置}$”是指依存关系中箭头的开始词(支配词)，例如，nsubj(赴京$_2$，柳生$_1$)表示“柳生”是“赴京”的名词主语。同理，图5例句中词之间的依存关系表示如下：

nsubj(赴京$_2$，柳生$_1$)
Root(Root$_0$，赴京$_2$)
dobj(赴京$_2$，赶考$_3$)
punct(赴京$_2$，，$_4$)
conj(赴京$_2$，行走$_5$)
nmod：prep(行走$_5$，大道$_{10}$)
case(大道$_{10}$，在$_6$)
nummod(大道$_{10}$，一$_7$)
nummod(一$_7$，条$_8$)
amod(大道$_{10}$，黄色$_9$)
case(大道$_{10}$，上$_{11}$)

分别统计小说、新闻及课本主题词的从属词，并按照它们之间依存关系的个数由高到低排序，结果如表6~8所示。

从表6发现，与小说主题词有关的从属词种类最多是所

属修饰关系(poss),涉及的从属词主要包括身体器官、亲属关系、内心活动、性格特征、社会角色等。经统计,与小说主题词相关的依存关系由高到低依次是nsubj、amod、dobj、poss,这些依存关系所对应的从属词主要是以小说主人公为核心而展开的多角度描写。

结合表7,以新闻的主题词“滑稽剧团”为例分析新闻语体的特征,与“滑稽剧团”有关的从属词主要是时间词(过去,目前、未来),地点词(上海、全国),描述其发展状态词(逐渐、缓慢、衰退),涉及的人主要有剧团的管理人员和演员等。由此可以看出,新闻是以叙述事件发生的时间tmod、地点及现状等为主的语体。

表7　新闻主题词的支配词分布

Tab. 7　Distribution of governing words of the topic words in news

关系名	数量	支配词
poss	242 109	衰败、上海、典型……
nsubj	118 764	进行、实现、发表、表示……
conj	85 952	和、及、同……
tmod	47 920	目前、当今、现在、未来……

课本由多种语体组合而成,其主题词的从属词分布如表8所示。这里以课本主题词“父亲”为例分析。“父亲”一词出现最多的是朱自清的散文《背影》。统计“父亲”有关的依存关系和与其对应的从属词,主要包括:nsubj(戴着、探身、穿过、笑、招手)、advmod(慢慢、蹒跚、挺拔)、poss(背影、皱纹、脸、身体、心)等。通过与“父亲”相关的从属词,可以感受到作者与父亲之间浓浓的父子之情。

表8　课本主题词的支配词分布

Tab. 8　Distribution of governing words of the topic words in textbook

关系名	数量	支配词
poss	14 567	无助、严重、认真、开心……
nsubj	11 761	显示、指出、认为、觉得、说明……
clmod	10 448	由于、因为、以为……
dep	3 811	这个、那、这些、那些、那人……

经过分析三种语体主题词的从属词及它们之间的依存关系可以看出,通过语义层面依存关系的挖掘使三种语体的本质特征已经显示出来了。此外,以上这些分析都是从词之间的关系出发所得到的,而词之间的搭配规则(句法结构)也是很重要的,接下来就从词之间的搭配规则出发,分析三种语体的差异。

4.1.3　主题词相关的句法结构分析

4.1.2节讨论的是与主题词相关的从属词及它们之间的依存关系,发现从这个角度出发,三种语体有较大的差异。本小节讨论的是与主题词搭配的规则(句法结构)有哪些,它们在不用的语体中是否有差异。首先,对主题词所在的句子建立句法树,以小说的主题词“柳生”为例,以句子“柳生 赴京 赶考 ,行走 在 一 条 黄色 大道 上 。”构建的句法树如图6所示。其次,找出与“柳生”有关的句法结构:IP→NP VP,NP→NN,VP→VP PU VP,VP→VSB,VSB→VV VV,VP→VV PP,PP→P LCP,LCP→NP LC,NP→QP ADJP NP,QP→CD CLP,CLP→M,ADJP→JJ,NP→NN。最后,统计全文跟“柳生”有较高相似度的句法结构并按降序排列。同理,对新闻、课本做相同的处理,得到与主题词有关的句法结构如表9所示。

从表9可以看出,与新闻主题词有关的句法结构最多,其次是小说,最后是课本。同样,以小说主题词“柳生”为例,与其有关的句法结构“IP →NP VV”,结合图5的句法树和图4的依存树,这个句法结构表明了“柳生”的动作是“赴京”,从依附“赴京”的支配词可以得知“柳生赴京”的目的是“赶考”。所以通过分析可以得到与小说语义主题词(“柳生”)相关的句法结构集及依存关系集,同时也可以得到与小说主题词(“柳生”)相关的核心动词集及依存词集,并对这些核心动词和依存词分别进行聚类,进而得到与小说主题词相关的核心动词块及依存词块。对于新闻和课本也采用同样的方法进行研究。

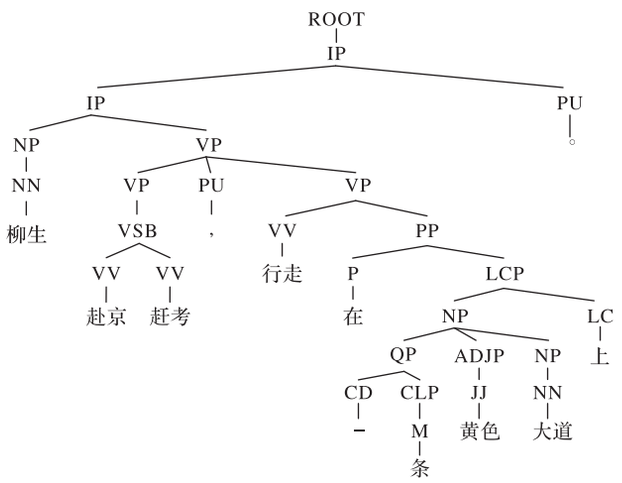


图6　句法树的例子

Fig. 6　Example of syntactic tree

表9　与主题词相关的句法结构(部分)

Tab. 9　Syntactic structure related to topic words (part)

语体	数量	句法结构
小说	189 389	IP →NP VV,NP → PN,VP → VV AS
新闻	208 938	NP → NN NN,NP → NR,NP → NT
课本	147 773	NP → DT NN,VP → VV VV,NP → PN

通过对小说、新闻及课本的语义主题词、依存关系及句法结构之间的内在联系进行分析,能让读者更加深刻地了解这三种语体每类特征之间的内在联系及它们所能反映的语体特征。

4.2　词的2元语体区分度分析

在作者识别任务中,词的N元能够很好地区分不同的作者,那么,在语体分类任务中,词的N元能否区分不同的语体。从表3的分类结果可以看出,词的2元对应的分类准确率较高,所以词的2元具有语体区分度。与词一样,词2元的频率分布与注意力分值及分类准确率的关系如表10所示。

表10　词2元的分值区间、频率及准确率的分布　单位:%

Tab. 10　Distribution of score interval, frequency and accuracy of bigrams of words　unit: %

注意力分值区间	频率比	准确率
[0.0,0.01)	79.38	46.25
[0.01,0.15)	13.98	74.65
[0.15,1.0]	6.64	91.88

从表10可以看出,用6.64%高注意力分值(大于等于0.15)的词的2元就能使分类准确率达到91.88%;而使用79.38%低注意力分值(小于等于0.01)的词的2元,对应的分类准确率是46.25%,这说明高注意分值的词的2元具有更好的语体区分度。通过训练词的2元,所得的高注意力分值的

词的2元如表11所示。

表11 高注意力分值的词的2元

Tab. 11 Bigrams of words with high attention score

语体	高注意力分值的词的2元
小说	鼠妹问、余占鳌司令、黄瞳道、福贵说、乡亲们、不知道、是吗、小屁孩、软骨头、站起来
新闻	据调查、刘代英坦言、贷款协议、记者追问、叶笃初表示、改革方案、报表范围、市场环境
课本	脚疼、特务们、喜儿说、父亲自言自语、四叔说、少爷好、或星期四、您吉祥、旧衬衣

从表11可以看出，在小说中，词的2元主要是“主语+动词”，例如：“鼠妹 问”“福贵 说”。经统计，小说中的动词多数是单音节，如“说”“喊”“问”。因为与双音节动词相比，单音节动词的动作性比较强，这充分体现了小说的另一面：以描写人物行为动作为主的语体。此外，小说中还有一些群体称呼（“乡亲 们”“姑娘 们”）及一些口语化的词或短语（“是 吗”“不 知道”），所以小说也具有口语的特征。新闻词的2元也是以“主语+动词”的结构为主，例如：“刘代英 坦言”“记者 追问”“叶笃初 表示”。与小说不同的是，这些动词大多数是双音节，所以这些动词比小说中的单音节动词更具有严谨性。例如：“表决”具有“说”的意思，但更多的是表示经过思考以后所做出的决定，其形式比较正式，这与新闻的特点相符。此外，新闻中还有VV+NN或NN+NN形式的词2元比较多，且这两个结构中的无论名词还是动词都倾向于双音节词。正如冯胜利所言，单双音节词具有语体区分度。由于课本包含多种语体形式，所以课本中的词的2元特点介于小说和新闻之间，其中小说部分类似于小说的特点，事实类文章类似于新闻。对于课本中其他的语体，本文暂不作讨论。

4.3 词类语体区分度分析

本文使用词类的含义见表12，词类的作用是指明词的性质，通过词类可用了解每一种语体关注的重点。词类在三个语体中的注意力平均分值如图7所示，从中可以看出：

1）三种语体的词类分值分布趋势相似，这说明每一种词类的语体区分度是相对比较稳定的。

2）从词类的分值大小来看，词类整体的分值都比较小，这说明词类具有较小的语体区分度。

3）词类的语体区分度由高到低依次是：ON、SB、IJ、LB、FW、MSP、DER、ETC、OD、BA、CS、DEV、VE、CC、SP、PN、VC、DEC、DT、JJ、NT、P、LC、AS、VA、CD、M、DEG、AD、VV。

从表3分类准确率来看，基于词类的分类准确率不高，且从图8可以看出，三种语体的注意力分值分布几乎重合在一起，这说明单纯词类特征并不能很好地区分小说、新闻及课本。所以，本文借助卡方检验来判断词类在三种语体中是否具有显著差异。

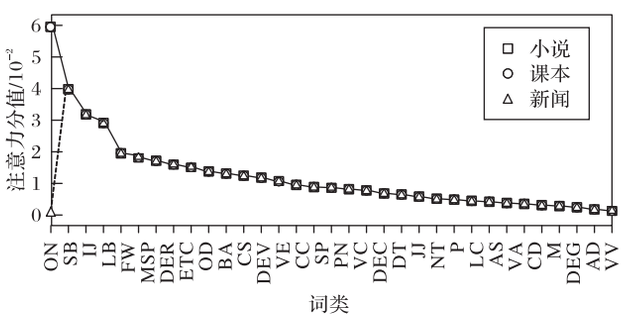


图7 词类的注意力分值分布

Fig. 7 Attention score distribution of POS

表12 宾州树库标记

Tab. 12 Symbols of Penn Treebank

标记	含义	标记	含义	标记	含义	标记	含义
AD	副词	DT	限定词，如“这”	SB	长“被”，“给”…	NT	时序词
AS	体态词/体标记，“着”	ETC	“等”，等等	SP	句尾小品词	DEV	“地”
BA	“把”和“将”	FW	外来词	VA	表语形容词	M	量词
CC	并列连词，如“和”	IJ	感叹词	VC	系动词，如：“是”	MSP	“所”
CD	数字，如：“一百”	NN	普通名词	VE	“有”	P	介词
CS	从属连词，如“如果”	LB	短“被”“给”…	VV	动词	PN	代词
DEC	“的”词性标记	LC	定位词，如：“里”	NN	普通名词	PU	标点
DEG(R)	联结词“的”/“得”	ON	拟声词，如：“哈哈”	NR	专有名词	OD	序数词

由于词类是离散型数据，且要检验它与三种语体的显著关系，故使用R x C列联表的卡方检验来验证，其原理跟卡方检验一样，是卡方检验的扩展。检验结果如表13所示，其中，卡方值按降序排列。在卡方检验中，特征的卡方值越大其在语体中就越显著，经过计算每一个词类的卡方值，最后得出32种词类在三种语体中都有差异，这里选择卡方值最大的NN（名词）进行分析，结果如表14所示。

4.4 词类的2元语体区分度分析

与词一样，词类的2元也具有语体区分度，词类的2元保留了比词类更多的词与词之间的共现信息。不同的词类2元平均注意力分值分布如图8所示。本节主要分析词类的2元，不包含标点（即PU标记）的词类2元。

从图8可以看出，具有语体区分度的词类的2元在三种语体中都是“NN + **”。从这三种语体词类的2元的数量来看，小说是20种，新闻是14种，课本是13种，即小说的2元结构最丰富，其次是新闻，最后是课本。从搭配词类的性质来看，小说中与NN搭配最显著是CD（数词），新闻中也是CD（数词），而课本是VA（形容词）。经统计发现，小说中的“NN CD”主要用于描述与人有关的特征，如“这娃20了”；而新闻中的“NN CD”主要描述一个事件相关的特征，如“滑稽剧团2012年开始衰退。”从这个角度来看，词类的2元（NN CD）可以看作小说和新闻的特征。另外，经统计发现数词在新闻中出现了29 121个、在小说中出现了6 826个，在课本中出现了2 378个，从这个角度来说，与CD搭配的词类的数量也存在着差异。对于词类的2元（NN VA）虽然在课本中较为显著，但是它在新闻和小说中也存在，例如：

小说：面色苍白、副官潇洒、高粱凄婉；

新闻：情况充实、特征明显、股市健康；

课本：人多、花朵大、榴莲贵、政策好；

在课本中，像“多、大、贵、好、热”等单音节形容词比较多，

其次是小说,最后是新闻,从这个角度来看,(NN VA)具有显著差异是合理的。

表 13　词类的卡方值分布

Tab. 13　Distribution of Chi-square value of parts of speech

序号	词类	卡方值	序号	词类	卡方值
1	NN	16 960.42	6	P	5 820.86
2	NT	16 078.02	7	ETC	5 372.11
3	CC	11 536.39	8	SP	3 828.25
4	JJ	7 693.97	9	OD	3 777.79
5	AS	7 105.36	10	DT	3 431.78

表 14　名词(NN)的卡方检验结果

Tab. 14　Results of Chi-square test of nouns

语体	名词	作用
小说	爷爷、老头子	表示人物称呼
	高粱、浪潮	表示事物名称
	心、眼睛、手	表示人体部位名称
	心里、思绪	表示抽象事物名称
	……	……
新闻	公司、集团、政府	表示机构名称
	人民、工人、教师	表示群体名称
	经济、政治、文化	表示宏观范畴名称
	银行、医院、法院	表示职能部门名称
	……	……
课本	社会、家、自己	表示人的社会关系名称
	思想、健康、理想	表示抽象事物名称
	榕树、流水,鸟	表示事物名称
	教室、操场、课堂	表示场所名称
	……	……

4.5　标点符号语体区分度分析

标点符号是书面语的有机组成部分,主要用来表示句子的停顿、说话者语气以及文本中词语的性质和作用。不同语体中标点符号的使用频率如图9所示。

从图9可以看出:逗号在小说中最多,其次是课本,最后是新闻;顿号在新闻中最多,其次是课本,最后是小说;引号在课本中最多,其次是小说,最后是新闻;感叹号在小说中最多、其次是课本、最后是新闻;问号同感叹号一样,都是小说中最多,其次是课本,最后是新闻。实验观察发现在新闻中,例如:"'冰棍论'、'靓女先嫁论'"这样的句子结构很多,通过顿号并列性质相同的词。引号主要出现在小说和课本的对话中,表示引出说话的内容;而在新闻中,引号主要用来表示一些具有特殊含义的人和物,例如:"房奴""寄生虫"等。最后,感叹号、问号、省略号这些带有情感色彩的标点,在小说和课本中更多。

图10给出了不同标点符号在不同语体中的平均注意力分值分布。可以明显观察到,省略号、问号、感叹号及冒号在三种语体中具有较大的语体区分性。由于标点符号的注意力分类准确率不高,与词类类似,本文利用卡方检验来检验标点符号在不同语体中的分布差异。

根据卡方检验的结果发现,省略号、感叹号、问号、顿号、句号、逗号、引号、破折号、冒号在三种语体中都具有显著差异,而分号在三种语体中的差异不明显。

接下来以最为显著的省略号为例,分析它在小说、新闻及课本中的分布差异。从数量上来说,小说中省略号出现了9 983次,新闻中出现了101次,课本中出现了2 188次。从省略号出现的场景来看,小说和课本中大约有80%的省略号都用于对话中,剩余的20%主要用于表示人物内心活动及用于列举内容的省略等场景中。而在新闻中,省略号主要用于列举内容的省略,避免啰嗦。接下来,通过具体的例子来分析,三种语体中常用省略号的例子如表15所示。从表15可以看出,小说和课本中的省略号赋予情感色彩,例如:小说中,"乡亲们接应我们来了,乡亲们来了……",这句话来自莫言的《红高粱》,讲述的是:面对日本侵略者的绞杀,在走投无路的情况下,余占鳌对豆官所说的话,体现出当时余占鳌看到来援救的相亲们所表现出的欣喜和激动。"我敬仰青松,但我却更喜欢榕树……"来自课本,选取秦似的《榕树的风度》。因为在原文这句话的前半句写了榕树的品质(榕树魁伟 、庄严、恬静、安祥),为了避免内容的重复,所以后面的省略号省略了作者喜欢榕树的原因。在新闻中,例句中的省略号省略了中国其他地方房价上涨情况,仅仅是列举内容的省略,不带有任何情感色彩。所以,从这个角度来看,省略号在三种语体中具有显著差异。

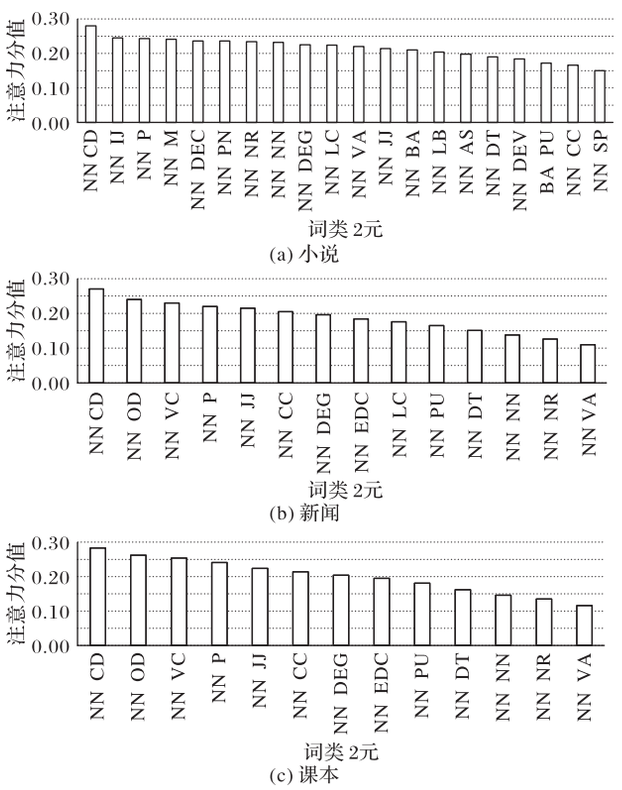


图8　词类2元的注意力分值分布

Fig. 8　Attention score distribution of bigrams of POS

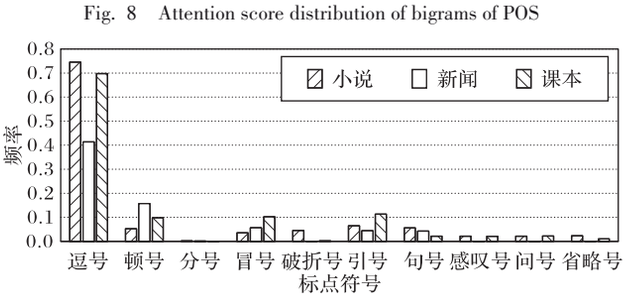


图9　标点符号的频率分布

Fig. 9　Frequency distribution of punctuations

4.6　标点符号2元语体区分度分析

本文忽略词,将连续出现的两个标点符号视为标点符号的2元,它能反映句子的结构和语气等信息,其注意力分值分

布如图11所示，其中横轴表示标点符号的2元序号，相应的对应关系如表16所示。

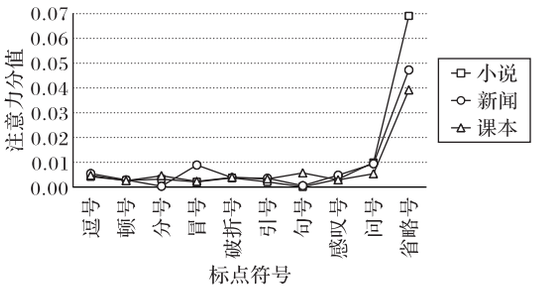


图10 标点符号的注意分值分布

Fig. 10 Attention score distribution of punctuations

表15 省略号在不同语体的例子

Tab. 15 Examples of ellipsis in different registers

语体	例句	含义
小说	乡亲们接应我们来了，乡亲们来了……	表示惊喜、兴奋
	广州中心城区房价上升了14%，大连	
新闻	上涨15.1%、呼和浩特上涨14.9%、南	用于列举的省略
	宁上涨9.9%、成都上涨9.2% ……	
课本	我敬仰青松，但我却更喜欢榕树 ……	省略喜欢榕树原因

从图11可以看出，小说中与省略号、感叹号、问号及引号组成的2元特征比较多。其中，标点符号的这些2元特征大部分来自人物对话或描述人物内心活动的句子。新闻中与省略号、感叹号及问号组成的2元特征主要用于对新闻事件相关的人物记录，属于客观陈述，不带任何情感色彩，这与新闻严谨的性质相符。而在课本中，这些2元特征以对话和内心感悟为主，进而引发学生思考，这与课本的目的相符。

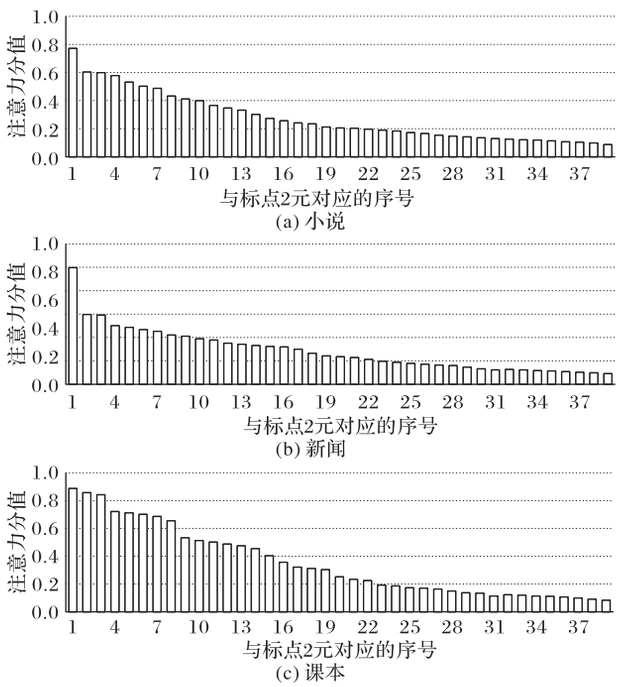


图11 标点符号2元的注意力分值分布

Fig. 11 Attention score distribution of bigrams of punctuations

表16 标点符号的2元映射示例说明

Tab. 16 Explanation of the example of binary mapping of punctuations

序号	图2中横轴序号对应的标点符号2元			序号	图2中横轴序号对应的标点符号2元			序号	图2中横轴序号对应的标点符号2元			序号	图2中横轴序号对应的标点符号2元		
	小说	新闻	课本		小说	新闻	课本		小说	新闻	课本		小说	新闻	课本
1	“……	……”	……”	11	—?	…”	》……	21	?!	、—	’……	31	“。	’“~	…”
2	‘……	)?	—。	12	=。	)]	:!	22	“…	-’	。”	32	‘’	’	‘’
3	)“	:)	》?	13	!?	—“	:!	23	“!	“。	——。	33	:”	:“	:)
4	(?	)!	:……	14	。”	”?	‘!	24	‘?	:。	“!	34	“。	、-	。。
5	“……	,!	“……	15	”……	,/	’!	25	…。	”。	、?	35	‘?	-”	》。
6	”……	!”	:。	16	”?	:?	?”	26	?!	,?	“。	36	”!	. :	:’
7	:……	:?	!”	17	?”	’-	)?	27	“…	~。	”?	37	“:	’--	—……
8	!”	:)	——!	18	‘。	]。	++	28	“!	》。	“?	38	”?	,:	)。
9	:?	’”	——?	19	:《	./	~!	29	“?	:。	)……	39	—。	-(	:——
10	:!	、?	:?	20	’!	-。	’?	30	’……	’……	’。				

通过以上各种统计和详细分析，本文将一元特征汇总如表17、18所示，其中：星号(*)表示该语体的显著特征，空白表示该特征在语体中不显著。这些主要是词、词类、标点符号、句法结构及依存关系。

2元特征是这些一元特征的组合，种类多且复杂，本文不再列出。

表17 语体的特征(词、词类)汇总

Tab. 17 Summary of stylistic features (words, POS)

语体	词															词类								
	人名词	亲属称呼词	地点词	专有名词	计量单位	单音节词	双音节词	重叠词	身体器官词	内心活动词	方言词	职能机构词	家庭角色词	社会角色词	外来词	时间词	人称代词	指示代词	语气词	感叹词	句尾小品词	瞬间动词	形式动词	
小说	*	*	*		*	*	*	*	*	*	*		*			*	*	*	*	*	*	*		
新闻	*		*	*	*		*					*		*	*	*							*	
课本	*			*			*					*		*	*	*	*	*		*	*		*	

表 18　语体的特征(标点符号、句法结构、依存关系)汇总

Tab. 18　Summary of stylistic features (punctuations, syntactic structures, dependency relationships)

语体	标点符号			句法结构								依存关系								
	疑问号	感叹号	省略号	NP→PN	NP→NN	NP→NR	VP→VV AS	VP→VV	VP→VV VV	NP→NT	NP→DT NN	名词主语	并列两个词	时间修饰词	直接宾语	语气词	形容词状语	依赖关系	所属修饰	因果从句
小说	*	**	*	**	**	*	*	*		*	*	*	*	*	*	*	*		*	*
新闻					*				*	*		*	*	*				*		
课本	*		*		*			*	*	*		*		*	*	*		*		*

5　结语

本文利用注意力网络模型提取能区分小说、新闻及课本的词、词类、标点符号、语法结构及它们的 $N(N=1,2)$ 元特征。相较其他三类特征,词汇特征更能直接反映出不同语体的区别,所以针对词汇特征,本文进行了深入分析(语义分析、依存关系和句法结构);对于词类和标点符号,由于注意力网络的分类准确率并不高,所以结合卡方检验一起分析。对于句法结构,借助句法树,将其序列化后,通过训练注意力网络挖掘出能区分不同语体的句法结构集。最后,通过多轮组合特征的训练,不但得到了每一种语体的关键特征集,而且还得出了每一种特征对不同语体的重要性。接下来将在以下几个方面进行改进工作:

1)提取能区分不同语体的其他特征。

2)分析影响注意力网络评分的因素,例如:句长,从而可以更好地完善模型。

3)改进注意力网络模型,将词在句子中的位置信息也考虑进来。

参考文献 (References)

[1] 霍小立. 语体特征及其影响变量的研究[D]. 南京:南京大学, 2014: 1-5. (HUO X L. The study on stylistic features and their influencing variables [D]. Nanjing: Nanjing University, 2014: 1-5.)

[2] 陶红印,刘娅琼. 从语体差异到语法差异(下)——以自然会话与影视对白中的把字句、被动结构、光杆动词句、否定反问句为例[J]. 当代修辞学, 2010(2): 22-27. (TAO H Y, LIU Y Q. From stylistic differences to grammatical differences (part two) — the cases of ba-repair, passive constructions, bare verb predicates, and negative interrogatives between natural conversation and media dialogue[J]. Contemporary Rhetoric, 2010(2):22-27.)

[3] 冯胜利. 语体语法的逻辑体系及语体特征的鉴定[J]. 汉语应用语言学研究, 2015(1): 1-21. (FENG S L. The logical system of stylistic grammar and the identification of stylistic features [J]. Research on Chinese Applied Linguistics, 2015(1): 1-21.)

[4] 张豫峰. "得"字句与语体的关系[J]. 河南大学学报(社会科学版), 2000, 40(1): 105-108. (ZHANG Y F. The relationship between "De" sentence and style[J]. Journal of Henan University (Social Science), 2000, 40(1):105-108.)

[5] 钱小飞. "地"字结构识别[J]. 现代语文(语言研究), 2006(5): 61-63. (QIAN X F. The "De" word structure recognition [J]. Modern Chinese, 2006(5): 61-63.)

[6] 方梅. 谈语体特征的句法表现[J]. 当代修辞学, 2013(2):9-16. (FANG M. Talking about the syntactic representation of stylistic features[J]. Contemporary Rhetoric, 2013(2): 9-16.)

[7] 林毓霞. 书面语体与标点符号[J]. 当代修辞学, 1987(3):11-12. (LIN Y X. Written style and punctuation[J]. Contemporary Rhetoric, 1987(3):11-12.)

[8] 胡骏飞,陶红印. 基于语料库的"弄"字句及物性研究[J]. 外语教学与研究, 2017, 49(1): 64-72. (HU J F, TAO H Y. A corpus-based study on the transitivity of "Nong" sentences and physical properties[J]. Foreign Language Teaching and Research, 2017, 49(1): 64-72.)

[9] 肖天久,刘颖.《红楼梦》词和 N 元文法分析[J]. 现代图书情报技术, 2015, 31(4): 50-57. (XIAO T J, LIU Y. An analysis of the words and N-grams for "A Dream of Red Mansions"[J]. New Technology of Library and Information Service, 2015, 31(4): 50-57.)

[10] 周浩. 基于神经网络的句法分析研究[D]. 南京:南京大学, 2017:30-53. (ZHOU H. Research on syntactic analysis based on neural network[D]. Nanjing: Nanjing University, 2017:30-53.)

[11] WANG Y, ZHANG J. Keyword extraction from online product reviews based on bi-directional LSTM recurrent neural network [C]// Proceedings of the 2017 IEEE International Conference on Industrial Engineering and Engineering Management. Piscataway: IEEE, 2017: 2241-2245.

[12] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2019-11-03]. https://arxiv. org/pdf/1409. 0473. pdf.

[13] PAPPAS N, POPESCU-BELIS A. Multilingual hierarchical attention networks for document classification[EB/OL]. [2019-11-03]. https://arxiv. org/pdf/1707. 00896. pdf.

[14] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[15] KALMAN B L, KWASNY S C. Why tanh: choosing a sigmoidal function [C]// Proceedings of the 1992 International Joint Conference on Neural Networks. Piscataway: IEEE, 1992: 578-581.

[16] SIDOROV G, VELASQUEZ F, STAMATATOS E, et al. Syntactic dependency-based n-grams as classification features [C]// Proceedings of the 11th Mexican International Conference on Artificial Intelligence, LNCS 7630. Berlin: Springer, 2012: 1-11.

This work is partially supported by the 2018 National Major Program of Philosophy and Social Science Fund (18ZDA238), the China's Ministry of Education Project of Humanities and Social Sciences (17YJAZH056), the Beijing Social Science Fund (16YYB021).

WU Haiyan, born in 1985, Ph. D. candidate. Her research interests include natural language processing.

LIU Ying, born in 1969, Ph. D., professor. Her research interests include corpus linguistics, computational linguistics, machine translation.