文章编号：

融合 CNN 和结构相似度计算的排比句识别及应用*

穆婉青[1]，廖健[1]，王素格[1,2]

（1. 山西大学计算机与信息技术学院，山西 太原 030006；

2. 山西大学计算智能与中文信息处理教育部重点实验室，山西 太原 030006）

摘要：排比句具有结构紧凑、句式整齐、富有表现力等鲜明的特点，广泛应用在各种文体之中，在近几年语文高考的鉴赏类问题中也多有考察，但在自动识别方面的研究还鲜有涉及。本文依据排比句结构相似、内容相关的特点，以句子的词性、词语作为基本特征，设计了融合卷积神经网络和结构相似度计算的排比句识别方法。首先将词向量和词性向量融入句子的分布式表示中，利用多个卷积核对其进行卷积操作，设计出基于卷积神经网络的排比句识别方法。利用分句之间的词性串构造相似度计算，设计了基于结构相似度计算的排比句识别方法。同时考虑句子内部的语义相关性和结构相似性，将卷积神经网络和结构相似度计算方法融合，用于排比句的识别。在文学作品数据集和高考题中的文学类阅读材料数据集进行排比句识别实验，验证了本文所提的方法是有效的。

关键词：排比句；语义相关性；结构相似性；卷积神经网络

中图分类号：TP391 **文献标识码：**A

A CNN and Structure Similarity Calculation Fused Method for Parallelism Recognition and Application

Mu Wanqing[1], Liao Jian[1], Wang Suge[1,2]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China; 2.Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract : Parallelism has the advantages of compact structure, neat sentence, expressiveness and other distinctive features in all kinds of literary forms. In recent years, parallelism has also been found as the problem of appreciation in the Chinese college entrance examination, but the research of automatic recognition is rarely involved. In this paper, according to the characteristics of the similar syntactic structure and content relevance in parallelism, make the part of speech and words become the basic characteristics of sentences, we design the fusion of convolutional neural network and structure similarity calculation method to recognition parallelism. We first add the word embedding and the vector of part of speech into the sentence distributed representation, using multiple convolution kernels to execute the convolution operation, design the parallelism recognition method based on convolutional neural network. Using the parts of speech of the clauses string to create similarity calculation, design the parallelism recognition method based on structure similarity calculation. Taking account of the semantic relevance and structure of the sentence similarity, we fusion the convolutional neural network and structure similarity calculation method to recognition parallelism. The experimental results show that the proposed recognition parallelism method is effective in the literature dataset and literature reading material datasets of the Chinese college entrance examination.

Key words: parallelism; semantic relevance; structure of the sentence similarity; convolutional neural network

---

* **收稿日期：** **定稿日期：**

基金项目： 国家“八六三”高技术项目（2015AA015407）；国家自然科学基金资助项目（61573231, 61632011, 61672331）；山西省科技基础条件平台计划项目(2015091001-0102)

作者简介：穆婉青（1993—），女，硕士研究生，研究方向为自然语言处理；廖健（1990—），男，博士研究生，研究方向为自然语言处理；王素格（1964—），女，教授，研究方向为自然语言处理。

1 引言

修辞格[1]是通过修饰、调整语句，运用特定的表达形式以提高语言表达作用的方式和方法。在文学类文本中恰当地使用修辞，可使语言表达更加的准确、生动，富有感染力。在众多修辞格中，排比作为一种使用频率较高的修辞格，受到了语言学家的广泛关注。典型的排比句[2,3]一般由三至五项组成，排比项[4]之间内容相关[5]、结构相似，词汇之间相互呼应，增强了文章的语篇连贯性，使行文丰满有力[6]，广泛的应用在政论、艺术[7]等语体中，在近十年的高考语文鉴赏题中也多有考察，以 2016 年北京市高考语文卷第 24 题为例。

问题：文章第四段运用了多种手法，表达了作者对老腔的感受。请结合具体语句加以赏析。

部分参考答案：排比。“这是……”、“亦或是……”、“也像是……”等句子使用了排比的手法，多角度表现老腔不同的腔调，充分表现老腔艺术魅力和我对老腔的沉迷。

若能自动识别文学类文本中的排比句，不仅可求解阅读理解内的鉴赏类问题，还可以进一步挖掘文本中的隐式情感，为计算机对文本的思想感情、语言风格等方面的自动赏析奠定基础。

排比句具有特征突出、节奏感强[5]，增强文章的表达效果的作用，引起不少语言学家的关注。高婉瑜[8]对对偶和排比两种辞格进行了详细的对比分析，认为二者仅在语法结构和语义上具有相似点，而在平仄相对、字面重复度等特征上不具有一致性。叶定国等人[9]对中文的对偶、排比辞格与英语的 Antithesis、Parallelism 辞格进行了分析对比，发现中文和英文的辞格并非完全对应，Antithesis 和 Parallelism 有包孕关系，而对偶和排比却相互独立，具有不用的语用功能。吕敬华[10]以《大堰河——我的保姆》为例，分析了排比在其中的艺术特色，认为使用排比使文章更有缠绵的韵味。何佳利[11]以《蜗居》的经典台词为例，研究了排比句在影视作品中作用，认为使用排比增强了语言气势，更好的表达了人物的思想感情。皮晨曦[7]研究了排比辞格在政论和艺术两种语体中的差异，认为排比在政论语体中常用来举例说明，而在艺术语体常用来抒发情感。张璐璐[12]分析了排比句在微博中的使用情况，认为使用排比使语言的表达更有感染力，并根据排比项的不同，将排比分为成分排比、分句排比、单句排比。张晓[13]从结构、词语重复度、排比项的数量出发对排比进行了再分类。上述文献仅仅从语言学层面对排比句进行了研究分析，而在自动识别方面的研究较少。梁社会等人[14]对《孟子》、《论语》中的排比句进行分析，联系排比句的结构、词语重复等特点设计了相应的排比句自动识别算法。该算法填补了对古汉语排比句自动识别的空白，但由于其只考虑了排比句结构类似和词语重复的表层特点，忽略了语义信息，且对语料本身有较高的依赖性，无法应用到现代汉语的语料中。

近年来，文本表示和句子分类方面的研究工作一直开展的如火如荼。相比之前经典的布尔模型和向量空间模型等文本表示模型，Word2vec[15-17]可以利用上下文语义信息对词语进行分布式表示，基本思想是通过 Skip-gram 和 Continuous Bag of Words（CBOW）模型将每个词映射成 n 维特征向量，通过词向量之间的距离获取词语之间的语义相似度，将其应用到文本的聚类、分类等自然语言处理领域的研究工作中。卷积神经网络[18,19]（Convolutional Neural Network，CNN）是一种前馈神经网络，由输入层、卷积层、池化层、全连接层、输出层组成，由于其良好的自学习能力和泛化能力，在短文本的表示和句子分类上也取得了一系列进展。Kim Y 等人[20]在预训练好的词向量的基础上，构造了一个双通道的 CNN 模型解决句子级别的分类任务，并取得了较好的效果。Kalchbrenner N 等人[21]提出了一种使用动态 pooling 方法的 Dynamic Convolutional Neural Network（DCNN）模型对句子语义进行建模，该模型在一定程度上保留了词序信息。Hu B 等人[22]也在 CNN 的基础上提出了一种解决句子建模的网络结构，主要用于解决句子匹配的问题。

本文依据排比句的定义[2-5]，概括了排比句的三个特点：（1）至少由三条相互衔接的排

比项组成；（2）各排比项之间具有语义相关性；（3）各排比项之间具有结构相似性。例如，下面的两个排比句。

例 1：烛光柔，月光静，电光更静，正如做事迅速的人，来去无声。

例 2：一粒种子，可以无声无息地在泥土里腐烂掉，也可以长成参天的大树；一块铀块，可以平庸无奇地在石头里沉睡下去，也可以产生惊天动地的力量；一个人，可以碌碌无为地在世上厮混日子，也可以让生命发出耀眼的光芒。

为了识别具有上述三个特点的排比句，本文结合文本的内容和结构，设计了基于 CNN 和结构相似度计算融合的排比识别方法，该方法不仅可以用于识别排比修辞，还可以为其他修辞格的自动识别、衡量文本的相似度、解答鉴赏类等问题提供参考。

2 排比句自动识别方法

对于排比句的自动识别，首先对一个句子进行结构化表示，然后对其是否为排比句进行判断。根据排比句中的排比项具有内容相关、结构相似的特点，本文以文本的词性、词语作为基本特征，同时考虑文本内部的语义相关性和结构相似性，设计了基于 CNN 和结构相似度计算的排比句融合识别方法。设 P_1、P_2 分别为利用 CNN 模型和结构相似度计算得到的该句子为排比句的概率，P_Y 是由 P_1 和 P_2 共同作用决定的排比句的最终概率，模型的框架图如图 1 所示：

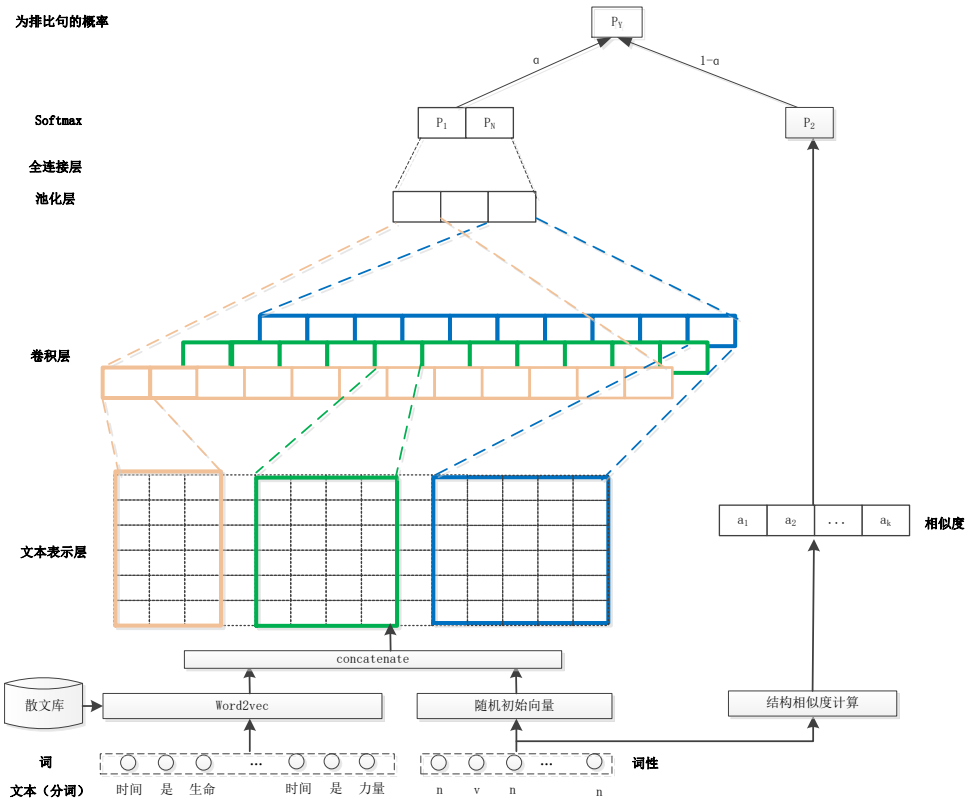

图 1 融合 CNN 和结构相似度计算的排比句识别框架图

图 1 中对文本进行分词等预处理操作后，以词语、词性为特征对文本内容进行分布式[23-25]表示，在此基础上，利用 CNN 对文本进行分类求得 P_1，利用 LCS 算法计算结构相似度求得 P_2，最终对 P_1、P_2 进行加权求和得到 P_Y。基于 CNN 排比句识别和基于结构相似度计算排比句识别详细介绍见第 2.1 节和 2.2 节。

2.1 基于 CNN 的排比句识别

（1）融合词语信息和词性信息的句子表示

由于文档中的句子是非结构化数据，需要将其转换成计算机可识别的结构化数据。Word2vec 是利用词语的上下文信息映射到 K 维向量空间上一种方法，它可使词语的语义表示信息更加丰富。本文利用 Word2vec 对大规模的数据进行训练获得词的分布式表示，在此基础上，加入句子的词性信息对句子进行表示。

设一个句子 S=($word_1, word_2, ..., word_n$) ，它是由 n 个词构成，其中第 i 个词 $word_i$ 的词性为 pos_i。利用 Word2vec 模型对大规模的散文数据进行训练后，$word_i$ 可以生成 kw 维的向量 xw_i。考虑到排比句具有结构相似的特点，本文视词性信息为浅层结构信息的刻画，将词语信息与词性信息共同作用对句子进行表示。假设现有的词类集为 C，对|C|中每一个词类 pos_i（i=1,2,…,|C|）进行编码，随机产生互不相同的维度为 kp 的向量 xp_i。对于每个词语与词性对<$word_i$, pos_i>（i=1,2,…,n）的联合表示，采用拼接技术生成一个 k 维的特征向量 X_i（i=1,2,…,n），这样融合了词语和词性两类信息特征的<$word_i$, pos_i>的表示见公式（1）。

$$X_i = \begin{pmatrix} xw_i \\ xp_i \end{pmatrix} \tag{1}$$

这里 X_i 为句子 S 中<$word_i$,pos_i>生成的特征向量，向量维度为 $k=kw+kp$，其中，kw 为向量维度，kp 为词性向量维度。对于未出现在 Word2vec 词表中的词语，其向量 xw_i 中的分量均以小于正整数 θ 的随机数进行填充。

设句子 S 表示的向量矩阵为 Sen^{kn}，融合语义信息和词序信息的句子 S 可由 n 个 X_i（i=1,2,…,n）进行拼接，其表示形式见公式（2）。

$$Sen^{kn} = \left[X_1 X_2 ... X_n\right] = \begin{bmatrix} xw_1 & xw_2 & ... & xw_n \\ xp_1 & xp_2 & ... & xp_n \end{bmatrix} \tag{2}$$

（2）基于多个卷积核的 CNN 的排比句识别

CNN 是一种前馈式神经网络，它利用反向传播算法对网络结构进行优化，对网络中的未知参数进行估计，常用于处理图像和时间序列等二维网格结构数据，容错能力强，运行速度快，具有较强的自适应能力，近期在句子表示和句子分类方面也获得不错的结果。本文利用大小不同的卷积窗口进行卷积操作，对获取的特征集合利用最大池化方法进行池化，为了防止训练样本过少造成的过拟合问题，在全连接层部分使用了 Dropout 技术，即以一定的概率暂时丢弃网络中的部分参数，将池化后获得的向量经过全连接层后，连接到 Softmax 层，最终获得该句子为排比句的概率。

设有 $t(t\epsilon\{1,2,...,n\})$个卷积核，$Sen^{kn}$ 为卷积操作输入层，$k\times c_m$ 为第 $m(m\epsilon\{1,...,t\})$个卷积核 C_m 的窗口大小，W^{kcm} 权重向量矩阵，w_j 为 W^{kcm} 中的第 $j(j\epsilon\{1,...,c_m\})$个权重向量，$f_a^{c_m}$ 为第 m 个卷积核获得的第 $a(a\epsilon\{1,...,n\text{-}c_m\text{+}1\})$个特征，$F^{c_m}$ 为特征集合，$f_{\max}^{c_m}$ 为 F^{c_m} 中的最大特征值，FW 经过卷积、池化、全连接操作后的特征向量，c_m 为卷积核每次卷积的词的个数。

基于 CNN 排比句识别算法（CNN）：

输入：句子 S 的表示 Sen^{kn}；

Step1 对 Sen^{kn}，利用 C_m 进行卷积操作，获得的第 a（$1\le a\le n\text{-}c_m\text{+}1$）个特征为 $f_a = \mathrm{f}(b + \sum_{a}^{c_m+a-1} w_i \cdot X_i)$；这里的“·”为两个向量做点积，$b$ 为偏置值。当卷积核的移动窗口

为 1 时，卷积后可获得长度为 $n\text{-}c_m+1$ 个特征集合 $\mathrm{F}^{c_m}=[f_1^{c_m}, f_2^{c_m},..., f_{n-c_m+1}^{c_m}]$；

Step2 利用 $t(1\le t\le n)$ 个卷积核分别对 Sen^{kn} 进行卷积操作，可获得 t 个特征集合 $\mathrm{F}^{c_1},\mathrm{F}^{c_2},...,\mathrm{F}^{c_t}$。

Step3 利用最大池化方法进一步减小特征，即对每一个特征集合 F^{c_m}，将最大值 $f_{\max}^{c_m}$ 作为输出 $f_{\max}^{c_m}=\max(f_1^{c_m}, f_2^{c_m},..., f_{n-c_m+1}^{c_m})$；

Step4 将池化层输出的多个特征集合的最大值 $f_{\max}^{c_1}$，$f_{\max}^{c_2}$，$\ldots, f_{\max}^{c_t}$ 进行全连接生成 FW；

Step5 将 FW 最终连接到 Softmax 分类器，输出 S 为排比句的概率 P_1；

Step6 当 $\mathrm{P}_1>\Phi_1$ 时，则句子 S 为排比句。

2.2 基于结构相似度计算的排比句识别

最长公共子序列（LCS，Longest Common Subsequence）[26]算法可计算出两个或多个序列中最长的公共子序列，该子序列中的元素在原序列中不一定是连续的，但是前后顺序不发生改变，广泛的用于计算图形、文字之间的相似度。本文依据 LCS 算法的特性，在句子词性串的基础上，利用该算法对句子的结构相似度进行度量。

由于排比句各排比项之间具有结构相似性的特点，本文以逗号和分号作为分隔符对句子进行分句，以连续的三个分句为一个分句单元，对各分句分词后生成词性串。在此基础上，利用 LCS 算法计算出两两分句之间的最长公共子序列；再利用相似度计算公式计算分句单元内的两两分句间结构相似度，并对其求平均值。最终将所有分句单元中句子间相似度平均值最高的认为是排比句。

设一个句子 $S=(s_1,s_2,...,s_t)$由 $t(t>2)$个分句组成，第 $i(i\in\{1,...,t\})$个分句 s_i 的词性串序列为 p_i，$\mathrm{len}(p_i)$为 p_i 的词性串长度，分句 s_i 和分句 $s_j(j\in\{1,...,t\})$的公共词性串为 p_{ij}、结构相似度值为 ss_{ij}，$\mathrm{sim}(s_i, s_j)$为相似度函数。假设连续的三个分句为一个分句单元，则第 $k(k\in\{1,...,t\text{-}2\})$个分句单元 SC_K 内为 a_k，所有分句单元中句子间相似度平均值最大为 a_{max}、$SC_K=\{s_k, s_{k+1}, s_{k+2}\}$。

基于结构相似度计算的排比句识别算法（SSC）：

输入：句子 S 中的分句单元 $SC_1,SC_2,\ldots, SC_{t-2}$;

Step1 对每一个单元 SC_K，执行 Step1.1-Step1.3

Step1.1 计算两分句 s_i、s_j 之间共同的词性串 $p_{ij}=h(s_i,s_j)=\mathrm{lcs}(p_i,p_j)$；

Step1.2 计算分句 s_i、s_j 结构相似度值[26]

$ss_{ij}=\mathrm{sim}(s_i,s_j)=(2\times\mathrm{len}(p_{ij}))/(\mathrm{len}(p_i)+\mathrm{len}(p_j))$；

Step1.3 计算 SC_K 的平均相似度值 $a_k=\dfrac{ss_{k(k+1)}+ss_{(k+1)(k+2)}+ss_{k(k+2)}}{3}$；

Step2 选取 t-2 个分句单元中的最大平均相似度值 $a_{max}=\max\limits_{k\in\{1,2,...,t-2\}}(a_k)$；

Step3 当 $a_{max}>\Phi_2$ 时，则句子 S 为排比句。

2.3 基于 CNN 和结构相似度计算的融合

排比句具有语义相关性，结构相似性的特点。由于 CNN 考虑了文本的语义信息，而 SSC 算法考虑了文本的结构信息，因此，将两者的结果进行融合，用于判别一个句子是否为排比句。为了融合结构信息，将第 2.2 节中获得的分句单元中最大平均相似度值 a_{max} 视作该句子为排比句的概率 $P_2=a_{max}$。设 α 为权重，利用 3.1 节求得的概率 P_1，判断句子 S 是否为排比句的最终概率为 $P_Y=\alpha P_1+(1-\alpha)P_2$，这里的 α 是依据排比句的标签信息确定的最优值，当 $P_Y>\Phi_3$ 时，则句子 S 为排比句。

3 实验及分析

3.1 数据集及评价指标

实验数据来源于高中语文课文、全国历年语文高考题的散文文本、查字典网（https://www.chazidian.com/）和散文吧网站（https://www.sanwen8.cn/），经过人工标注的排比句 2000 条，非排比句 2000 条，共计 4000 条，其中非排比句来源于排比句的上下文。为了训练 Word2vec 模型，采用 1946-2006 年近 3.5G 的《人民日报》语料、散文吧网站上获取的散文 71460 篇、全国历年语文高考题的散文文本 184 篇。所有的实验数据均经过去噪、去重、分词等预处理操作。

实验结果采用的评价指标为精确率（precision）、召回率（recall）、F1 值和正确率（accuracy）。

3.2 参数设置

使用第 2.1 节 CNN 模型对排比句识别时，训练过程中参数的更新采用随机梯度下降方法。通过多次实验，选取性能最优参数如下：$kw=200$，$kp=200$，$t=3$，C_1、C_2、C_3 的卷积窗口大小分别为 300×3、300×4、300×5，学习率为 0.01，dropout 的概率为 0.5，$\Phi_1=0.5$，$\Phi_2=0.65$，$\Phi_3=0.6$，$\alpha=0.3$。

实验结果均采用五折交叉验证，即每次取全部数据的 80%为训练集，其余 20%为测试集，各评价指标均重复五次实验取平均值。

3.3 实验结果及分析

实验 1：不同权重 α 下的融合 CNN 和 SSC 排比句识别结果

为了分析权重 α 对第 2.3 节提出的融合 CNN 和 SSC 方法判别排比句的影响，本实验选取 $\alpha\epsilon\{0.1,0.2,...,0.8,0.9\}$，进行了 9 组对比实验，实验结果如图 2 所示，其中，横坐标为权重 α 的取值，纵坐标为实验结果。

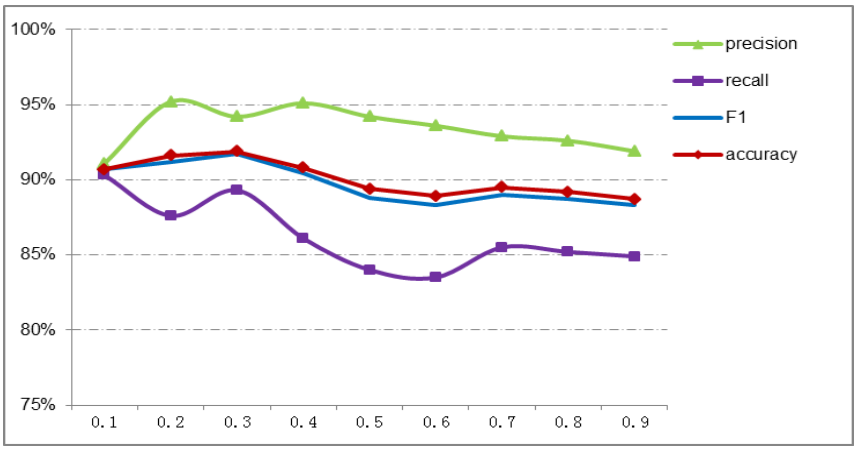

图 2 不同的权重 α 下融合 CNN 和 SSC 的排比句识别结果

由图 2 可看出：随着 α 的变化，各评价指标均有所波动，说明内容和结构对于排比句识别均有一定的影响。当 α 值为 0.3 时，F1 值和 accuracy 取得了最优结果，说明结构相似度

（SSC 方法）在排比句识别起着重要的作用。在后续实验中均选取 α=0.3。

实验 2：不同特征、分类方法进行排比句识别比较

为了验证本文提出方法的有效性，我们设置了如下对比实验：

SVM[w]：以 $word_i$ 为特征的基于 SVM 分类器的排比句识别；

SVM[pos]：以 pos_i 为特征的基于 SVM 分类器的排比句识别；

SVM[w+pos]：以 $word_i$ 和 pos_i 为特征的基于 SVM 分类器的排比句识别；

CNN[w]：仅以 $word_i$ 为输入特征，利用 2.1 节提出的基于 CNN 的排比句识别方法；

CNN[pos]：仅以 pos_i 为输入特征，利用 2.1 节提出的基于 CNN 的排比句识别方法；

CNN[w+pos]：利用 2.1 节提出的基于 CNN 的排比句识别方法；

SSC[w]：利用 2.2 节基于 SSC 的排比句识别方法，仅以 $word_i$ 为输入特征，且令 $\mathrm{sim}(s_i, s_j)$为 cosine 相似度 $\mathrm{sim}_{\cos}(s_i, s_j)$，计算内容相似度；

SSC [pos]：利用 2.2 节基于结构相似度计算的排比句识别方法；

SSC [w+pos]：利用 2.2 节基于结构相似度计算的排比句识别方法，以 $word_i$ 和 pos_i 为输入特征，且令 $\mathrm{sim}(s_i, s_j)=\beta\mathrm{sim}_{\mathrm{lcs}}(s_i, s_j)+(1-\beta)\,\mathrm{sim}_{\cos}(s_i, s_j)$，计算内容相似度和结构相似度，其中 β 取最优值为 0.6。

CNN[w+pos]+ SSC [pos]：利用本文第 2.3 节提出的排比句识别方法。

各分类方法的实验结果如表 1 所示：

表 1 各类排比句识别方法的实验结果比较

评价指标 / 方法	precision	recall	F1	accuracy
SVM[w]	71.1%	63.9%	67.2%	68.2%
SVM[pos]	73.6%	64.0%	68.4%	70.4%
SVM[w+pos]	75.3%	68.4%	71.6%	72.9%
CNN[w]	89.3%	84.7%	87.0%	87.2%
CNN[pos]	88.3%	80.2%	84.1%	84.7%
CNN[w+pos]	89.5%	85.9%	87.7%	87.6%
SSC[w]	95.6%	68.1%	79.5%	82.5%
SSC[pos]	90.9%	86.0%	88.4%	88.7%
SSC[w+pos]	96.0%	80.0%	87.3%	88.3%
CNN[w+pos]+SSC[pos]	94.2%	89.3%	91.7%	91.9%

由表 1 可以看出：

（1）从分类器的角度，在三种不同的特征表示下，三种分类方法识别排比句的结果分别为 SVM[w]<SSC[w]<CNN[w] 、 SVM[pos]<CNN[pos]<SSC[pos] 、 SVM[w+pos]<SSC[w+pos]< CNN[w+pos]，由此，SVM 识别排比句的效果是三种方法中最差的，主要原因是利用 SVM 忽略了句子的序列信息，不能对句子的深层的非线性语义特征进行建模。

（2）对于 CNN 算法，在 precision，F1 和 accuracy 三种评价指标下，CNN[pos] <CNN[w] CNN[w+pos]，CNN[w+pos]是识别排比句的效果最好，而 recall 指标下，CNN[pos]< CNN[w+pos]<CNN[w]，CNN[w]是识别排比句的效果最好，说明 CNN 模型可以有效利用句子中词语信息。

（3）对于 SSC 算法，在 precision，recall，F1 和 accuracy 四种评价指标下，SSC[w] <SSC[w+pos]<SSC[pos]，SSC[pos]是识别排比句的效果最好，说明 SSC 模型可以有效利用句子中结构信息。

（4）综合词语与词性作为特征的角度，SVM[w+pos]<SSC[w+pos]<CNN[w+pos]，说明

利用词语信息与词性信息共同对句子进行表示，可以弥补单一特征表示的不足，也证明了我们所采用的特征的有效性。

（5）从方法融合的角度，CNN[w+pos]+SSC[pos]所有方法中识别排比句的效果最好的，说明融合后的方法充分的考虑了句子的内容相关性和结构相似信息。

对于第 1 节中例 1 和例 2，依据 CNN[w+pos]、SSC[pos]、CNN[w+pos]+SSC[pos]的概率如表 2 所示。

表 2 方法融合前后的概率值对比

示例	CNN[w+pos]	SSC[pos]	CNN[w+pos]+SSC[pos]
例 1	0.85	0.53	0.63
例 2	0.34	0.84	0.69

由表 2 可知，例 1 中使用 SSC[pos] 得到排比句的概率值为 0.53，判断为非排比句。例 2 中使用 CNN[w+pos]得到的概率值为 0.34，也判断为非排比句。然而，两个例子若使用融合后的方法的概率值分别为 0.63 和 0.69，均可正确的判断为排比句。由此，对于排比句来说，有些句子侧重结构上的相似，有些句子侧重语义上的相关，将两个方法融合是必要的。

实验 3 CNN[w+pos]+SSC[pos]在高考鉴赏题中的应用实验

利用验证本文所提的方法应用在高考题解答中的作用，选取 2017，2016 和 2007 年北京高考题中文学类阅读材料进行排比句识别，获得排比句识别的概率见表 3 所示。

表 3 文学类文本中排比句识别的概率值

文本来源	例句	概率值
2017 年	那些将天边画出蜿蜒起伏线条的山丘，那些怒放成海洋或孤零零独自开放的鲜花，那些低头吃草或昂头沉思的马群……	0.692
2016 年	这是发自雄浑的关中大地深处的声响，抑或是渭水波浪的涛声，也像是骤雨拍击无边秋禾的啸响……	0.601
2007 年	那风里雨里，透明的阳光里，透明的流水里，有我湿湿的想念，永远永远。	0.714

由表 2 可看出，三个例句的概率值均大于 Φ_3，则对排比句识别均正确，因此可知本文提出的方法可以有效的对排比句进行识别，进而为高考阅读理解内的鉴赏题解答和文本的自动赏析服务。

4. 总结

本文针对文学类文本中的排比句，依据其内容相关、结构相似的特点，设计了融合 CNN 和结构相似度计算的排比句识别方法，并与其他方法做了比较实验。针对高考题内的文学类阅读理解的问题，将本文所提方法应用到阅读材料中的排比句识别中，取得了不错的效果，从而进一步证明了本文所提出的方法和特征选择的有效性。但本文仍有不足，例如数据量较小，没有在政论、新闻等其它语体上进行对比实验。今后的工作可以考虑如何更好的将文本的结构相似的信息嵌入到 CNN 的输入层中，并将其应用到更为广泛的语体语料之中。

备注：为了便于更多的研究者开展排比句识别方面的研究，我们将本文所用的数据共享到山西大学文本情感分析技术资源开放与服务平台（http://115.24.12.5/）。

参考文献

[1] 张宗正. 修辞格位、修辞格变体和修辞格作品——关于修辞格本质即同一性的再思考[J]. 修辞学习，2003, (2): 24-25.

[2] 李胜梅. 排比的篇章特点[J]. 南昌大学学报(人文社会科学版)，2005，36(5)：127-133.

[3] 陈永敬. 排比的构成特征及排比项数限制的心理机制[D]. 武汉：华中师范大学, 2008.
[4] 夏丽芳. 排比的语篇衔接功能[J]. 牡丹江教育学院学报，2008，(1)：51-52+85.
[5] 范俊. 排比的语用修辞研究[D]. 重庆：四川外国语大学，2013.
[6] 聂仁海. 排比辞格—语言本质的典型体现[J]. 现代企业教育，2006，(9)：160-161.
[7] 皮晨曦. 政论语体与艺术语体排比差异研究[D]. 广州：暨南大学，2011.
[8] 高婉瑜. 谈对偶与排比[J]. 修辞学习，2008，(5)：58-60.
[9] 叶定国. Parallelism 与对偶、排比[J]. 外语与外语教学，1999，(2)：53-55.
[10] 吕敬华. 排比和反复的妙用[J]. 长沙电力学院学报(社会科学版)，2000, 15(2)：96-98.
[11] 何佳利. 《蜗居》比喻和排比辞格探析[J]. 大众文艺，2011，(4)：125-126.
[12] 张璐璐. 浅析微博语言中的排比修辞格[J]. 青春岁月，2012，(24)：120-121.
[13] 张晓. 排比的再探讨[J]. 昭乌达蒙族师专学报(汉文哲学社会科学版)，2004，25(8)：17-18+11.
[14] 梁社会, 陈小荷, 刘浏. 先秦汉语排比句自动识别研究——以《孟子》《论语》中的排比句自动识别为例[J]. 计算机工程与应用，2013，49(19)：222-226.
[15] Mikolov T，Sutskever I，Chen K，et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the International Conference on Neural Information Processing Systems. USA：Curran Associates Inc，2013：3111-3119.
[16] Mikolov T，Chen K，Corrado G，et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv：1301. 3781, 2014.
[17] Le Q V，Mikolov T. Distributed Representations of Sentences and Documents[J]. arXiv preprint arXiv:1405.4053，2014
[18] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 6(40)：1-23.
[19] 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述[J]. 计算机应用, 2016, 39(09)：2508-2515+2565.
[20] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv：1408.5882，2014.
[21] Kalchbrenner N，Grefenstette E，Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv：1404.2188，2014.
[22] Hu B，Lu Z，Li H，et al. Convolutional neural network architectures for matching natural language sentences[C]// Proceedings of Advances in Neural Information Processing Systems. USA：MIT Press，2014：2042-2050.
[23] Bengio Y，Schwenk H，Senécal J，et al. Neural probabilistic language models[J]. Journal of Machine Learning Research，2003，3(6)：1137-1155.
[24] Collobert R，Weston J，Bottou L，et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research，2011，12(1)：2493-2537.
[25] Kiros R,Zhu Y，Salakhutdinov R，et al. Skip-thought vectors[C]// Proceedings of International Conference on Neural Information Processing Systems. USA：MIT Press，2015：3294-3302.
[26] Lin C Y，Och F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics，USA：Stroudsburg ，2004.

作者联系方式：
作者一： 穆婉青，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，030006，18234129508，948611255@qq.com；
作者二：廖健，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，

030006,15935619813，liaojian_iter@163.com；
作者三：王素格，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，
030006,13934649855，wsg@sxu.edu.cn。