文章编号：1003-0077(2019)01-0077-08

基于篇章修辞结构的自动文摘连贯性研究

刘 凯，王红玲

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：尽管抽取式自动文摘方法是目前自动文摘领域的主流方法，并且取得了长足的进步，但抽取式自动文摘形成的摘要由于缺乏句子之间的合理指代或篇章结构，使得文摘缺乏连贯性而影响可读性。为提高自动摘要的可读性，该文尝试将篇章修辞结构信息应用于中文自动文摘。首先，基于汉语篇章修辞结构抽取摘要，然后使用基于LSTM的方法对文本连贯性进行建模，并使用该模型对文摘的连贯性做出评价。实验结果表明：在摘要抽取方面，基于篇章修辞结构的自动文摘相比于传统的抽取方法具有更好的ROUGE评价值；在使用基于LSTM连贯性模型评价摘要连贯性方面，篇章结构信息在自动抽取文摘时可以很好地提炼出文章的主旨，同时使摘要具有更好的结果。

关键词：篇章修辞结构；中文自动文摘；连贯性；可读性；实体网格模型；LSTM

中图分类号：TP391 **文献标识码**：A

Research on Automatic Summarization Coherence Based on Discourse Rhetoric Structure

LIU Kai, WANG Hongling

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: In order to improve the readability of automatic summaries, this article attempts to apply the discourse rhetorical structure information to Chinese automatic summarization. First, abstracts are extracted based on the rhetorical structure of Chinese texts. Then the LSTM-based methods are adopted to evaluate the coherence of the abstracts. The experimental results show that, automatic abstraction based on discourse rhetorical structure has better ROUGE value than traditional methods. The coherence evaluation results show that the discourse structure information can help the system extract the subject of the article automatically.

Keywords: discourse rhetoric structure; Chinese automatic summarization; coherence; readability; entity-grid model; LSTM

0 引言

随着大数据时代的来临，用户时刻都在接收海量的信息，这使得信息冗余的情况经常发生，产生信息过载的问题。如何从大量的信息中找出有效信息是大数据时代面临的一个挑战，自动文摘技术则是应对该项挑战的利器。自动文摘是指通过自动分析给定的一篇或多篇文档，提炼、总结其中的要点信息，最终形成一篇长度较短、可读性良好的摘要。简言之，文摘就是通过对原文本进行压缩、提炼，为用户提供简明扼要的文字描述。

自动文摘是自然语言处理（natural language processing，NLP）的一个非常重要的领域，已经被研究多年，也涌现出许多方法，如基于语言分析的方法、基于统计的方法、基于聚类的方法和基于图的方法等，在部分自动文摘问题的研究上取得了明显的进展，并成功地将自动文摘技术应用于搜索引擎、新闻阅读等产品与服务中。但是自动文摘技术还远谈不上完美，特别是在摘要的可读性上还面临相当多的挑战和难题。其中，目前主流的抽取式自动文摘[1]表现较好，但从语言学角度上看往往不尽如人

收稿日期：2018-04-03 **定稿日期**：2018-06-26

基金项目：国家自然科学基金(61402314)

意,造成这种结果的原因很多,例如,提取的文章特征不能够很好地表达文章的含义,选取的句子不是文章的主要部分等,其中比较核心的问题是从文本中抽取出来的摘要句,它们之间的指代关系和篇章结构没有很好地被保留,这使得文摘不连贯,导致摘要的可读性不好,如"但就整个世界经济而言,其他国家的强劲增长势头会弥补这一损失。报告估计1997年世界经济增长百分之三点二,预计1998年将增长百分之三。"两句话之间并无直接的关系,而且由于抽取第一句还缺少成分(即"这一损失"的具体指向),导致这段摘要的质量不高,这也是抽取式摘要目前所面临的主要困难。

在语言学中,篇章(discourse)是由一系列连续的词、短语、子句或段落构成的语言整体单位[2]。文档摘要的实质也是篇章,提高摘要的质量,可从篇章分析着手。从篇章的角度考虑,一段语篇是否具有较好的质量主要从篇章的7个基本特征[3]来看,分别是衔接性(cohesion)、连贯性(coherence)、意图性(intentionality)、可接受性(acceptability)、信息性(informativity)、情景性(situationality)和跨篇章性(intertextualinty)。篇章的衔接性和连贯性,是篇章表层的形式表示,而连贯性[4]作为衡量篇章可读性的一个指标,表示各个句子之间有一定的顺序,句子的上下文之间有一定的承接,这对于多句文本在句法和逻辑上有着重要的意义。

一篇高质量的自动摘要,不仅需要确保能够最大限度地表达原文的含义,还要保证其在描述上前后一致、表达连贯,即具有良好的连贯性,以使读者具有良好的阅读感观。因此,本文尝试使用深层语言信息——篇章修辞结构,进行抽取式自动文摘的研究,重点考虑修辞结构在自动文摘选择核心内容时的影响,并基于篇章连贯性评价方法对抽取出的摘要进行连贯性评价。

1 相关研究

近年来,自然语言处理的研究对象逐渐从词汇、句法等浅层语义,深入到句子、篇章的语义连贯性和结构衔接性等深层语义方面。篇章修辞结构就是指句子之间或篇章之间的主次关系。从理论上说,通过分析篇章修辞结构,不仅能够抽取出篇章的主要信息作为摘要,还可以使文摘的语义更加连贯,从而提高自动摘要的质量。

从以往文献来看,基于篇章修辞结构的自动文摘在英语上的应用相对较多;而在汉语上,由于缺乏篇章修辞结构的标注语料,对该方面在自动文摘的应用上尚未见到。Marcu[5]从语言学的角度分析了篇章修辞结构信息中的核心作为摘要的原理,并基于RST标注语料构建了一个自动文摘系统。Yoshida等[6]将基于依存结构树的自动摘要看作一个树背包问题,提出了一种新的篇章结构,把依存结构转化为修辞结构,并就转换时出现的问题进行了改进。Louis等[7]分别基于RST和PDTB语料来说明篇章的结构信息和语义信息对文摘内容选择的影响,并且与非篇章特征(位置、句子长度等)进行比较。实验结果表明,内容选择的主体主要还是依靠结构信息,语义信息可以作为结构信息的一个补充。Goyal等[8]利用RST标注语料,将结构信息具体应用在内容的选择上,他们提出了一种新的监督学习的方法——SampleRank,通过在RST树形态转变的时候赋予不同的权值来计算每个篇章最小单元(EDU)的得分,最终选出得分最高的单元作为文摘,结果证明这样的方法是有效的。Mithun等人[9]基于博客文本构建了一个修辞结构的语料库,并利用篇章的结构特征抽取句子,结果表明,抽出的句子具有一定的连贯性。这进一步说明,篇章结构的特点不仅能够选择有代表性的内容,还能够使选出的内容具有一定的连贯性。

文摘连贯性研究是篇章连贯性的研究内容之一,主要研究摘要中句子与句子之间的连贯程度。当前篇章连贯性建模的主要工作分为三大类:局部篇章连贯性模型、全局篇章连贯性模型和混合篇章连贯性模型,这是依据语篇的跨度来进行划分的。其中局部篇章连贯性模型研究取得了相对较好的实验性能,代表性模型有:基于实体的模型(Barzilay和Lapata[10-11])、基于篇章关系的模型(Louis和Nenkova[12],Lin,et al[13])和基于神经网络的模型(Li[14],Nguyen[15],林睿[16],Xu等[17])。

当前的连贯性评价方法中,实体网格方法是一种用来评价局部连贯性的常见模型,它最早由Barzilay等[11]提出,依据衔接性理论,从句子间的表层连接上对连贯性进行建模,利用句子间相似名词实体的概率转移来为篇章的连贯性打分。

虽然Barzilay的方法是一种适用范围较广的方法,但对于一些特殊的语料来说,加上其他的特征则会对此方法有较好的提升。Strube等[18]给实体网

格方法添加了相关实体的语义信息，使用一个标注好的德文语料，在原先实体网格方法的基础上添加了实体的语义相关性(GermaNet API)，具体方法为当一个新的实体出现时，首先计算它和实体集合中实体的相关性，如大于某个值(t)，则将先前实体的信息(句法信息)分配给它。

基于神经网络的模型中，Li 等[14]在基于神经网络的基础上分别使用循环神经网络(recurrent neural network)和递归神经网络(recursive neural network)来对句子进行向量化处理，直接从语义的角度来考虑篇章的连贯性。两者虽然都是树模型，可 Recurrent 模型是将句子中每个单词的信息向后进行累积，并以最后一个单词的向量来表示整个句子的含义。Nguyen 等[15]继承了实体网格的特点，将神经网络方法和实体网格法结合，该文总结了实体网格方法的缺点——当窗口相当大的时候，可能会发生维数灾难，即计算得到的连贯性的值是一个很小的值。而该文给出的方法是用卷积神经网络将事先处理好的实体网格进行卷积操作，并把一个篇章的实体网格的内容映射为特征向量，再使用 Softmax 函数来对篇章的连贯性求解值，最终，通过比较 ROUGE 值发现，他的结果比 Li 等[14]的模型效果好，提高约 8 个百分点。在中文方面，Xu 等[17]的 recursive 模型，在 ROUGE 值方面略有提升。

2　基于篇章修辞结构的抽取式摘要

2.1　语料介绍

苏州大学自然处理实验室在针对连接依存树和篇章结构分析研究的基础上，吸取了修辞结构理论的树形结构和篇章主次关系，参考了宾州篇章树库对连接词的处理方式，同时结合汉语复句和句群理论，提出了基于连接依存树的篇章结构理论。本文使用的汉语篇章结构树库(Chinese discourse tree bank，CDTB)[19]就是基于该理论所标注的一个篇章修辞结构语料库，它结合了 RST(Rhetorical Structure Theory)[20]和 PDTB(Penn Discourse Tree Bank)[21]的特点，不仅具有篇章的结构信息，同时还具有语义信息。孙静等[22]利用最大熵等方法标注了 CDTB 中因果类、并列类、解说类和转折类的隐式关系，并与 PDTB 进行了对比，表明 CDTB 具有篇章语义信息。李艳翠等[23]分析了连接词在篇章中的作用：不同的连接词表示不同的篇章层次，从而表现出不同篇章单元的主次关系，以及连接词的隐显关系代表了篇章单元不同的语义信息。

该语料分别标注篇章的宏观修辞结构和微观修辞结构，其中宏观表示段与段之间的关系，微观表示段内的篇章单元的修辞关系，二者都是通过连接词将相应的篇章单元连接起来。

2.2　摘要抽取方法

2.2.1　抽取段落摘要

CDTB 采用树的形式表示汉语的篇章结构，包含了两种树的结构，其一是段内的篇章修辞结构树，即为每一个段落构建一棵篇章修辞结构树，其结构如图 1 中的例子所示，其中叶子节点表示一个具体

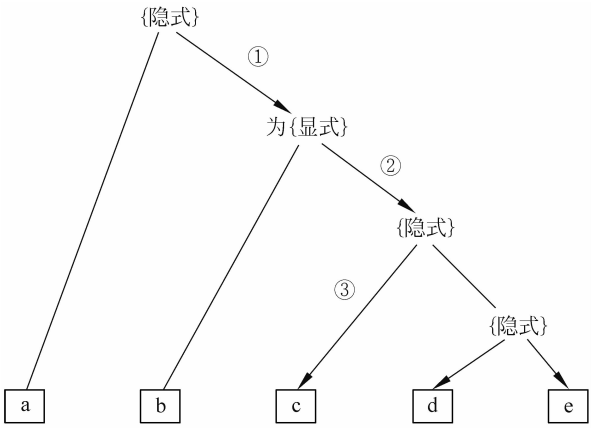

图 1　段内关系篇章结构示例图

的篇章基本结构单元，即 EDU(element discourse unit)，内部节点表示的是连接词，这里的连接词是指篇章连接词，它连接各个篇章单位，而根据连接词是否在篇章单位中出现又分为隐式和显式，总共标注有 657 篇；其二是段间的篇章修辞结构树，即段与段之间的篇章关系。

具体例子如图 2 所示，其中内部节点表示段间的关系，包括并列、转折、因果、解说四个大类，这几个大类又分为 17 个小类；不同的关系类型有不同的侧重点，根据它们的重要性可以确定关系主次，如转折关系中往往后一部分较为重要，因此树中的箭头将指向后一部分，目前总共标注了 97 篇。另外，为方便评价，我们对这 97 篇文档标注了人工摘要，包括文档中每段的段落摘要和整篇文档的摘要。

图 1 表示段内的各个子句的树形结构，图 2 表示段落间的树形结构，而 a～e 为子句，具体表示为：

a：如今，甘肃省的外资企业已不再为投资风险担忧。

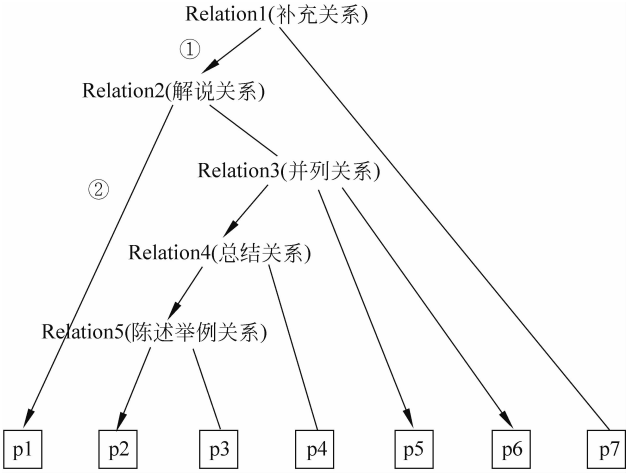

图 2 段与段关系篇章结构示例图

b：为确保对外开放的顺利进行。

c："八五"期间(一九九一至一九九五年)，甘肃省涉外保险业发展迅速。

d：累计已经提供六百一十一亿元的风险保障。

e：承保范围包括财产、责任、信用、人身四大类主、副六十多个险种。

通过对微观篇章修辞结构主次的标注，可以清晰地从图中观察到篇章的主要部分。为了更好地对各个子句进行重要性的计算，本文引入子句中词语的统计信息和层次信息，选择候选的摘要句。如式(1)所示。

$$\varphi = \sum_{i=1}^{N} \frac{S_i \times \frac{1}{M} \times \sum_{j=1}^{M} \mathrm{TFIDF}(x_{i,j})}{\mathrm{Depth}(i)} \quad (1)$$

其中，S_i 表示该子句是否为主要部分，即如图 1 中①②③所指示的路径，若是主要部分则为 1，否则为 0；M 表示该子句词语的个数；TFIDF 表示计算 tf-idf 值的方法；$x_{i,j}$ 表示子句 i 句的第 j 个词；Depth 表示子句 i 所在的层次。

具体的算法如下：

算法 1：
Input：一棵建好的段落篇章结构树
Output：抽取出的段落摘要
获取树中的一个节点
If 这个节点是一个叶子节点 then
　　依据标注信息中的 center 信息和公式(1)对各个叶子节点打分，选择候选摘要，center 中'3'表示各个篇章单位的地位相同，"1"表示前一个篇章单位更重要，"2"表示后一个篇章单位更重要
Else if 该节点有一个孩子节点 then
　　判断这个孩子节点的内容是否居于主要地位，如果是则继续迭代；否则依据公式(1)计算得分
Else 该节点有两个或以上的孩子节点 then
　　依据 center 的值选取居于主要地位的孩子节点，继续进行迭代。
End

2.2.2 抽取全局摘要

抽取全局摘要的方法和抽段内摘要的方法类似，依据标注的宏观的篇章修辞结构信息，得出在全局中重要的段落，并将主要段落的候选摘要作为全局的候选摘要。如图 2 所表明的顺序①、②所示，第一段位置是全文摘要产生的地方，也就是将第一段的候选摘要作为全文的候选摘要，最后利用候选摘要句得分和全局摘要长度的限制生成全局的摘要，本文抽取的全局摘要长度根据人工摘要的长度确定，平均长度为 80 个字。

由于 CDTB 语料库没有对单个子句作为一个段落进行处理，这样导致了标注的语料中漏掉了一些段落，因此这一步作为抽取段落摘要的补充，将源语料中漏掉的部分作为单独的一个段落摘要加入抽取完的摘要集合中，使得段落摘要的顺序和原始语料的段落对应。

2.3 基于 LSTM 的连贯性评价

由于循环神经网络能够有效地依据时间关系来对模型进行整合，通过输入的训练数据到循环的各个隐藏层的映射，能够学习到输入序列随着时间变化而变化的语义，而 LSTM 更适合处理相当长的序列，以及预测时间延迟非常长的重要事件。因此，本文首先采用 LSTM 来构建句子的分布式向量表示形式。

参考 Xu 等[17]引入实体连接对篇章连贯性的重要作用，本文也在模型中加入实体的向量，区别在于本文实体向量并不进行句子向量间加减的操作，而

是直接加在句子向量的后面，以希望保留更多原有的句子中的信息。具体模型图如图 3 所示。

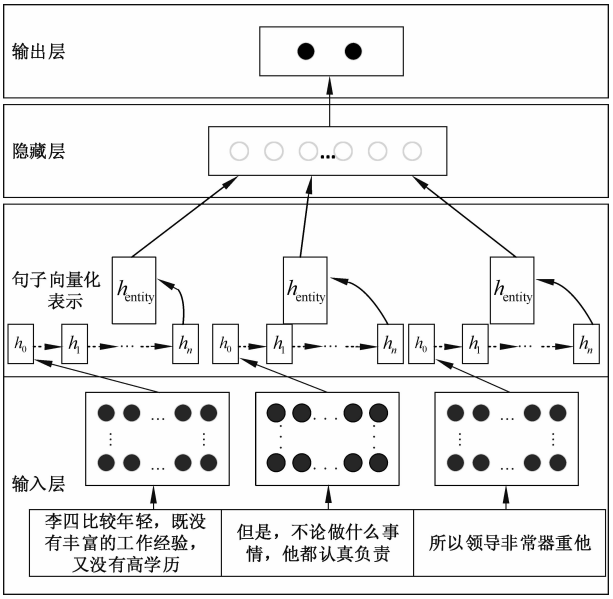

图 3　篇章连贯性模型图

由图 3 可知，该模型的组成分为三个部分，分别是分布式的句子向量、合成的句子和实体的向量以及基于团块的篇章连贯性表示。

为了判断这个输入的内容是否连贯，本文在训练时把计算篇章的连贯性看作一个分类任务，即对文本的内容进行排序，其中原始文本的内容为正例，经打乱后的文本为负例，并以此进行输入。

首先，它的输入是一个经过分词的句子块(也称作团)，表示这个篇章中的句子，可以是正例或负例。然后，将它用预先训练好的向量表示输入到一个 LSTM 层中，得到各个句子的向量。由于最后的全连接层输入要求是有相同长度的向量，而每个句子中的实体个数并不相同，因此必须进行一些额外的处理，使得实体向量的长度相同，本文是将把句子中的所有实体向量之和作为实体驱动向量。最后，将一个团块的实体驱动向量输入到最终的分类器中进行计算，判断该团块中的句子是否连贯。

2.3.1　实体驱动的句子向量表示

词向量是一种语言模型训练得到的产品，它表示词语在当前语言模型中所在空间的位置，即它形象地表示了词语之间的相互关系，而且它还会从语言模型的语义空间中迁移出一些未知的知识，这对进一步表示句子有相当大的帮助。因此，本文首先使用 glove[①] 工具训练出语料中的词向量，每个词向量的大小为 w。输入的句子可以看成一个词向量的组合，若用 x_i 表示句子 S 中第 i 个词向量，则长度为 N 的句子可以被表示为 $S_N = [x_1, x_2, \cdots, x_N]$，其大小为 $N \times w$，最终通过 LSTM 训练得到 k 维的句子向量 S_N，其大小为 $1 \times k$。

借助于前文对实体网格的分析中，我们发现实体的转移概率是判断一个篇章是否连贯的一项重要参数，它反映了人们在阅读时习惯于记忆的相同或语义相近的名词，以增加对文章内容的理解。因此，我们需要将实体向量与句子向量相融合。

一般而言，每个句子中的实体数量不同，因此就需要对实体向量进行归一化处理，本文采用将实体向量相加的方法，得到最终实体向量的表示，并将它直接拼接在句子向量之后。

2.3.2　句子间连贯性判断的团块表示

由于不同长度的篇章难以一起训练，所以给篇章规定了一个固定长度(滑动窗口)以方便计算，在此设滑动的窗口的长度为 C。将人工摘要按逗号隔开，经过统计分析，得到的平均句子长度是 2.6，因此选取 C 的值为 3 进行训练。

一个团块是由长度为 C 的连续句子组成的篇章，为了计算它的连贯性，将团块向量化，并输入到模型中。模型的公式如式(2)、式(3)所示。

$$q_c = f(\boldsymbol{W}_{\text{sen}} \times [\boldsymbol{h}_c, \boldsymbol{h}_{\text{entity}}] + \boldsymbol{b}_{\text{sen}}) \tag{2}$$

$$P(y_c = 1) = \text{sigmoid}(\boldsymbol{U}^{\text{T}} \cdot q_c + b) \tag{3}$$

其中 $[h_c, h_{\text{entity}}]$ 表示将实体向量拼接在句向量后面，q_c 表示隐藏层的输出，$\boldsymbol{W}_{\text{sen}}$ 表示句子向量，b_{sen} 表示偏置，$\boldsymbol{U}^{\text{T}}$ 表示大小为 $1 \times H$ 的向量，其中 H 表示隐藏层神经元的个数，最终 P 表示连贯的可能性，将之视为连贯性的得分。对于篇章 d 的连贯性，则将篇章中所有团块的连贯性得分相乘，即 $S_d = \prod_{c \in d} P(y_c = 1)$。

2.3.3　模型训练和优化

在训练时，我们将连贯性的计算看成分类的问题，把打乱顺序的文本看成是负例，原始文本看成为正例。在利用训练集和验证集数据进行模型学习之后，用该模型对测试集打分得到文本的连贯性，最后再利用最优参数对摘要连贯性进行打分。

由于是分类任务，因此该模型采用广泛应用的交叉熵函数作为目标函数，如式(4)所示。

① http://nlp.stanford.edu/projects/glove

$$J(\Theta)=\frac{1}{M}\sum_{c\in \text{trainset}}\{-y_c\log[p(y_c)]-(1-y_c)\log[1-p(y_c=1)]\}+\frac{Q}{2M}\sum_{\theta\in\Theta}\theta^2 \tag{4}$$

其中，Θ是模型需要训练的所有参数，M表示训练集的大小，Q是正则化项，防止模型的过拟合。

反向传播的梯度更新操作如式(5)所示。

$$\theta_T=\theta_{T-1}-\frac{\alpha}{\sum_{t=0}^{T}\sqrt{g_t^{i^2}}}g_t^i \tag{5}$$

g_t^i表示参数θ_i的第t步反向传播的梯度。在优化损失函数$J(\Theta)$方面，由于Adagrad非常适合处理稀疏数据，是应用范围较广的优化算法之一，本文也采用了Adagrad方法。每次迭代过程中，每个参数优化时使用不同的学习率，对于出现频率较高的参数采用较小的步长更新。

2.4 实验设置和结果分析

2.4.1 实验设置

(1) 基于篇章修辞结构的摘要抽取

抽取摘要所用的语料来自于CDTB，共97篇，平均每篇大约有600字。人工摘要平均约80字，因此抽取出大约80字的机器摘要。具体的抽取方法参见2.2节的描述。

(2) 连贯性评价

连贯性评价实验的语料来自于CTB9.0，从chtb_0001到chtb_1 151，以及chtb_2 000值chtb_3 145，共2 296篇语料中选出2 000篇作为此次的实验语料，这些当中包含了新闻、杂志等不同类型的文章。然后利用这些文章随机生成20个打乱顺序的文本，作为负类。

将语料分为训练集、验证集和测试集，所占比例分别为60%、20%、20%。训练时，为了防止过拟合问题，设置dropout rate为0.4，滑动窗口的大小为3。为了记录每次迭代后最好的值，在每训练100轮数据时将最好的模型保留下来，并且为了防止训练过程长时间无更新权值操作，设计当超过1 000轮未更新时，说明参数已收敛，很难继续学习，因此将强制终止学习过程。设置Adagrad的初始学习速率为0.001，并将词向量的维度设置为100。

为说明基于LSTM方法的有效性，本文采用基于实体网格的方法来与该方法的结果进行对比，并对结果进行分析。实验结果的评价指标为文本正确分类的准确率。

2.4.2 结果分析

(1) 摘要抽取评价结果

利用ROUGE方法，最后得到了基于篇章修辞结构的段落摘要和全局摘要的得分，其中表1为段落摘要的得分，表2为全局摘要的得分，表3为不同摘要抽取方法的对比实验结果。

表1 段落摘要评价结果

	Rouge-1	Rouge-2	Rouge-L
P	0.587	0.346	0.330
R	0.914	0.726	0.793
F	0.695	0.433	0.337

表2 全局摘要评价结果

	Rouge-1	Rouge-2	Rouge-L
P	0.909	0.687	0.637
R	0.795	0.551	0.473
F	0.833	0.572	0.450

表3 不同抽取方法对比实验结果

	Rouge-2(*F*值)	Rouge-L(*F*值)
段落摘要	0.433	0.337
全局摘要	0.572	0.450
TFIDF	0.417	0.318
LexRank	0.506	0.433
Lsa	0.443	0.366
PageRank	0.492	0.391

从表1中可看出，基于篇章修辞结构抽取出来的段落摘要的召回率R比较高，说明这种方法在抽取段落摘要时比较有效，同时由于段落摘要是相对较短的文本，所以它的准确率和F值较低；而全局摘要包含了所有重要段落的内容，因此准确率比较高(表2)。

为了对比不同抽取方法，我们分别使用基于图的方法中的PageRank、LexRank，基于统计方法的TF-IDF，基于聚类方法中的LSA等方法来与基于修辞结构的自动文摘进行对比，表3显示了这些方法的结果。

对比结果显示，基于篇章修辞结构的全局摘要抽取方法在Rouge-2和Rouge-L的F值上结果最好，而这两个值分别代表的是二元和最大长度子串的匹配程度。这说明了利用篇章结构来抽取文摘，

能够较好地保持文摘中长句子的准确程度，更能准确表达文章的含义。

（2）连贯性评价结果

由于目前缺乏通用的连贯性评价方法，本文采用基于实体网格的方法来与基于 LSTM 方法的结果进行对比，并对结果进行分析。实验结果的评价指标为将文本正确分类的准确率，得到结果如表 4 所示。

表 4　文本排序的评价结果

连贯性模型	准确率
Li[14] Recurrent 模型	0.708
Li[14] Recursive 模型	0.715
Fan Xu[17] 模型	0.724
基于实体的模型	0.623
基于 LSTM 的模型	0.986

可以看出基于实体的模型由于提取的特征较为简单，在分类时错分的概率比较大，而基于神经网络的模型，由于大量数据的输入和提取到的更丰富的特征，能够更加准确地分类出连贯的文本，具有更好的可靠性。

具体的示例如下：

① TF-IDF 摘要：浙江省今后将进一步提高对外开放水平，把全面推进对外开放向高层次、宽领域、纵深化发展作为重点。实施出口商品“龙头”计划，引导外资投资方向，探索新的投资方式。

② 篇章修辞摘要：浙江省今后将进一步提高对外开放水平，努力扩大对外贸易、利用外资和国际经济技术合作，并逐步完善对外经贸营销网络。

③ 人工摘要：浙江省今后将进一步提高对外开放水平，努力扩大对外贸易、利用外资和国际经济技术合作，并逐步完善对外经贸营销网络。把全面推进对外开放向高层次、宽领域、纵深化发展作为重点。实施出口商品“龙头”计划，引导外资投资方向，探索新的投资方式

通过对以上的示例可以看出人工摘要具有更好的连贯性和信息性，而篇章修辞结构和 TF-IDF 方法抽取的摘要概括了文章的信息，但仔细观察后可以看出篇章修辞方法抽取的摘要具有更好的连贯性。

在此基础上，本文利用 LSTM 模型评价了抽取摘要的连贯性，得到结果如表 5 所示。从表 5 可看出，相比于常见的摘要抽取方法，基于篇章修辞结构抽取出的摘要在连贯性评价方面具有较高的取值，这是由于篇章的主次结构是自然表述的过程，这使得按照篇章修辞结构抽取出的句子往往带有一定的连贯性，提升了抽取摘要的质量。同样地，我们也可以看出，人工摘要具有最高的连贯性得分，这一方面验证了人工摘要具有很强的连贯性，另一方面也说明基于篇章修辞结构的自动文摘在连贯性方面还有提升的空间。

表 5　摘要连贯性的评价结果

摘要的抽取方法	连贯性取值
篇章修辞摘要	0.573
TF-IDF 摘要	0.532
PageRank	0.503
Lsa	0.486
LexRank	0.562
人工篇章	0.858

3　结束语

本文使用抽取式的方法，利用汉语篇章结构树库中标注的篇章修辞结构信息，抽取出文档的重要部分作为文档摘要，并使用 ROUGE 和连贯性评价的方法分别对摘要的信息覆盖度和连贯性进行评分。实验结果表明，基于篇章修辞结构的文摘方法在这两个评分标准上都具有较好的性能。

一篇文摘质量的高低不仅仅在于内容的选择，它的表达形式（连贯性和衔接性）也相当重要。对于依据篇章修辞结构信息抽取得到的文摘来说，尽管由于篇章结构关系本身具有一定的连贯性，但相比与人工摘要还有很大差距，因此如何使用深层语义信息（如篇章话题结构）优化文摘连贯性将成为下一步的研究目标。同时，未来将使用最新研发的 discourse parser 对大规模自动文摘语料标注篇章修辞结构，然后再应用本方法抽取摘要。

参考文献

[1] Mani Inderjeet, Mark T Maybury. Advances in automatic text summarization[M]. Cambridge, MA: MIT Press, 1999.

[2] 徐凡，朱巧明，周国栋. 篇章分析技术综述[J]. 中文信息学报，2013，27(3)：20-33.

[3] de Beaugrande R. Dressler W. Introduction to text linguistics[M]. London,UK: Longman,1981.
[4] 殷习芳,刘明东.语篇连贯性研究综述[J].湖南第一师范学报,2006(3): 124-127.
[5] Marcu D. Discourse trees are good indicators of importance in text[J]. Advances in Automatic Text Summarization,1999: 123-136.
[6] Yoshida Y, et al. Dependency-based discourse parser for single-document summarization[C]//Proceedings of the 2014 EMNLP,2014: 1834-1839.
[7] Louis A,Joshi A,Nenkova A. Discourse indicators for content selection in summarization[C]//Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics,2010: 147-156.
[8] Goyal N,Eisenstein J. A joint model of rhetorical discourse structure and summarization[C]// Proceedings of the Workshop on Structured Prediction for NIP, 2016: 25-34.
[9] Mithun S,Kosseim L. Discourse structures to reduce discourse incoherence in Blog summarization[C]//Proceedings of the RANLP,2011: 479-486.
[10] Barzilay R,Lapata M. Modeling local coherence: An entity-based approach[C]//Proceedings of the 43rd ACL,2005: 141-148.
[11] Barzilay R,Lapata M. Modeling local coherence: An entity-based approach[J]. Computational Linguistics, 2008,34(1): 1-34.
[12] Louis A, Nenkova A. A coherence model based on syntactic patterns[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,2012: 1157-1168.
[13] Lin Z,Ng H T,Kan M Y. Automatically evaluating text coherence using discourse relations[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,2011,1: 997-1006.
[14] Li J,Hovy E. A model of coherence based on distributed sentence representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,2014: 2039-2048.
[15] Nguyen D T,Joty S. A neural local coherence model [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1320-1330.
[16] 林睿. 基于神经网络的篇章一致性建模[D]. 哈尔滨: 哈尔滨工业大学硕士学位论文,2015.
[17] Xu F,et al. An entity-driven recursive neural network model for Chinese discourse coherence modeling[J]. arXiv preprint arXiv: 2017.8201,2017.
[18] Strube M,Strube M. Extending the entity-grid coherence model to semantically related entities[C]//Proceedings of the 11th European Workshop on Natural Language Generation. Association for Computational Linguistics,2007: 139-142.
[19] 李艳翠. 汉语篇章结构表示体系及资源构建研究[D]. 苏州: 苏州大学博士学位论文,2015.
[20] Mann W C, Thompson S A. Rhetorical structure theroy: Toward a functional theory of text organization[J]. Text,1988,8(3): 243-281.
[21] Prasad R,et al. The Penn Discourse TreeBank 2.0 [C]//Proceedings of the 6th International Conference on Language Resourses and Evalution,2008,24(1): 2961-2968.
[22] 孙静,等. 汉语隐式篇章关系识别[J]. 北京大学学报(自然科学版),2014,50(1): 111-117.
[23] 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.

刘凯(1992—),通信作者,硕士研究生,主要研究领域为自然语言处理。
E-mail: 2274710776@qq.com

王红玲(1977—),副教授,主要研究领域为自然语言处理。
E-mail: hlwang@suda.edu.cn