基于主题相似度的宏观篇章主次关系识别方法*

蒋峰，褚晓敏，徐昇，李培峰，朱巧明

（苏州大学计算机科学与技术学院，江苏 苏州 215006；

江苏省计算机信息技术处理重点实验室，江苏 苏州 215006）

摘要：篇章分析是自然语言处理领域的一个重要任务。分析篇章主次关系有助于理解篇章的结构和语义，并为自然语言处理的应用提供有力的支持。本文在微观篇章主次关系识别研究的基础上，重点研究宏观篇章主次关系，提出了一种基于 word2vec 和 LDA 的主题相似度的宏观篇章主次关系识别模型。基于 word2vec 的主题相似度和基于 LDA 的主题相似度在不同维度上计算语义相似度，两者在语义层面形成互补，因而增强了模型识别宏观篇章主次关系的能力。该模型在宏观汉语篇章树库（MCDTB）上实验的 F1 值达到 79.9%，正确率达到 81.82%，相较基准系统分别提升了 1.7%和 1.81%。

关键词：宏观篇章主次关系；主题相似度；word2vec；LDA

中图分类号：TP391　　**文献标识码：**A

A Macro Discourse Primary and Secondary Relation Recognition Method Based on Topic Similarity

Jiang Feng, Chu Xiaomin, Xu Sheng, Li Peifeng, Zhu Qiaoming

(School of Computer Sciences and Technology, Soochow University, Suzhou, Jiangsu 215006, China; Provincial Key Laboratory for Computer Information Processing Technology, Suzhou, Jiangsu 215006, China)

Abstract: Discourse analysis is an important task in the field of Natural Language Processing. The analysis of discourse-level primary and secondary relations helps to understand the discourse structure and semantics, and provides strong support for the applications of Natural Language Processing. Based on the research of micro discourse-level primary and secondary relation recognition, this paper aims at macro discourse-level primary and secondary relation and provides a recognition model based on topic similarity with word2vec and LDA. The topic similarity based on word2vce and the topic similarity based on LDA calculate the semantic similarity on different dimensions. They are complementary at the semantic level, which enhances the ability of the model to recognize the macro discourse-level primary and secondary relations. Experimental results on the Macro Chinese Discourse TreeBank (MCDTB) show that our model achieves the F1-score of 79.9% and the accuracy of 81.82%, which improves the baseline by 1.7% and 1.81%, respectively.

Key words: macro discourse-level primary and secondary relation; topic similarity; word2vec; LDA

1 引言

近年来，自然语言处理领域的研究内容，逐步从浅层次的词汇、句法分析延伸到了深层次的语义理解。因此，自然语言处理研究的文本颗粒度，从单个词、短语、句子，延伸至句群、段落、篇章。篇章分析是目前研究的热点和重点，其目的是进一步研究自然语言文本的内在结构并理解文本单元间的语义关系，挖掘出文本的结构化和语义信息。

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（No. 61272260）；教育部中国移动科研基金（No. MCM20150602）；江苏省科技计划（No. SBK2015022101）

作者简介：蒋峰（1994——），男，硕士研究生，主要研究方向：自然语言处理、篇章分析；李培峰（1971——），男，教授，硕士生导师，主要研究方向：中文信息处理、事件抽取；朱巧明（1963——），男，教授，博士生导师，主要研究方向：中文信息处理，Web 信息处理。

篇章主次关系表示了篇章内部或篇章与篇章间的主要内容和次要内容的关系。其中，主要内容是指篇章中居于支配地位、起决定作用的部分，而次要内容是指篇章中居于辅助地位、不起决定作用的部分[1]。篇章主次关系主要分为微观和宏观两个层面，微观主次关系是指篇章中的一个句子内部的主次关系或两个连续句子间的主次关系，而宏观主次关系则是更高层次的主次关系，表现为段落、章节间的主次关系。研究篇章主次关系，有助于更好地认识和理解篇章的中心主题，更有效地挖掘篇章的宏观主题和篇章各部分之间的语义关联，并为自然语言处理的相关应用，如信息抽取[2] 、自动文摘[3] 、问答系统[4] 等提供支撑。

本文以 CTB8.0 中一个篇章（chtb_0056.nw.raw）为例来说明宏观篇章主次关系，如例 1 所示（完整的宏观篇章关系结构标注如图 1 所示）。在图 1 所示的树形结构中，自然段落是叶子节点（如段落 *a*、*b* 等），篇章关系为非叶子节点（如 *R*2、*R*3 等），箭头指向篇章关系中较为重要的部分。本文将篇章主次关系分为 3 类：1）P-S（Primary and Secondary），即主要在前，次要在后；2）S-P（Secondary and Primary），即主要在后，次要在前；3）M-P（Multi-Primary），即前后同等重要。

中国高新技术开发区发展迅速成果显著

a) 新华社北京十二月十七日电（记者秦杰）中国五十三个国家高新技术开发区发展迅速，已形成一大批机制灵活、适应市场经济要求、技术创新能力强的高新技术企业。

b) 中国高新技术开发区酝酿于八十年代初。到去年为止，中国高新技术开发区技术工贸年总收入达二千三百亿元，利税总额达二百三十八亿元，年出口创汇达四十三亿美元，均比创办初增长数十倍。其中，形成了一批具有一定规模的高新技术支柱产业，产值超亿元的企业达四百零五家，产值超十亿元的大企业四十二家。

c) 一九九六年，中国高新技术开发区企业研究开发投入达六十二点三五亿元，占企业产品销售收入的百分之三点五，开发、生产高新技术产品一万三千多种。

d) 近年来，中国高新技术开发区初步建立了适应社会主义市场经济体制和高新技术产业发展需要，与国际惯例接轨的管理体制和运行机制，建立并不断完善了包括信息、金融、法律、资产评估、产权交易等中介和服务机构，初步形成了适于高新技术产业发展的较为完善的支撑服务体系。

e) 为规范高新区的管理，依法治区，中国颁布了《国家高新技术产业开发区管理暂行办法》，同时长春、苏州、沈阳、长沙、石家庄、昆明等高新区也先后完成了高新区的人大立法工作或以政府令形式发布了高新区管理办法。（完）

例 1 chtb_0056.nw.raw 文章内容

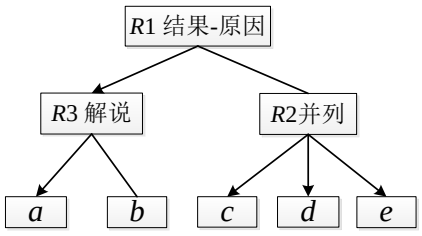

图 1 宏观篇章结构的树形表示（chtb_0056.nw.raw）

该例中，段落 *a* 中提出中国高新技术开发区发展迅速，段落 *b* 是对中国高新技术开发区发展情况的详细介绍。因此，主要是段落 *a* 的内容与外界发生语义关系，段落 *a* 主要，段落 *b* 次要，而且段落 *a*、*b* 间形成了解说关系；而段落 *c*、*d*、*e* 分别从 3 个方面阐述了中国高新技术开发区发展迅速的原因，因此三者为同等重要，形成并列关系。

本文组织结构如下：第 2 部分从理论、语料、模型三个方面介绍了篇章主次分析的相关研究工作；第 3 部分介绍了宏观汉语篇章语料库的建设；第 4 部分给出了一个基于主题相似

度的宏观篇章主次关系识别框架，并介绍了计算主题相似度的算法；第 5 部分详细分析了实验结果；第 6 部分总结全文并指出下一步工作。

2 相关工作

理论研究方面，在微观篇章关系上，Mann 和 Thompson[5-6] 的修辞结构理论（RST）根据修辞关系提出了“核-卫星”（Nucleus- Satellite）模式，并将篇章关系分为单核关系和多核关系两大类。对于单核关系来说，有关系的两个篇章单元一方为核心，一方为卫星。对于多核关系来讲，篇章关系连接的两个篇章单元同等重要，没有主次之分。在宏观篇章关系上，Van Dijk[7] 的篇章宏观结构理论提出了篇章宏观结构，宏观结构与微观结构相对，是篇章整体上的高层次的结构，每一层的宏观单元都由下一层的宏观单元通过归总形成，代表更为主要的篇章内容。

目前涉及到篇章主次关系语料资源主要包括修辞结构篇章树库（RST-DT）和汉语篇章树库（CDTB）等。修辞结构篇章树库（RST-DT）是以修辞结构理论（RST）为支撑，标注了篇章单元、篇章关系、主次关系（即“核心”和“卫星”）和篇章结构等，从而生成有层次的篇章结构树。汉语篇章树库（CDTB）是基于连接依存树的篇章结构理论，在宾州大学汉语树库（CTB）上标注了 500 篇微观篇章关系结构，共计 2342 个段落。该语料库是在每个段落上，自顶向下的标注一棵篇章关系结构树，其篇章基本单元为子句。RST-DT 和 CDTB 都进行了微观篇章主次关系的标注，但均未进行宏观篇章主次关系的标注。

微观篇章主次关系方面的计算模型研究较为广泛。在修辞结构篇章树库（RST-DT）上，Hernault[8] 使用的是开源的 HILDA 分析器， HILDA 分析器使用两个支持向量机（SVM）来进行构建篇章树，其分析器在篇章主次关系识别任务中的 F1 性能为 61.3%。Joty[9] 在他们前期[10] 句内篇章结构分析的工作基础上，分别应用句内和句间两个动态条件随机场模型（DCRF），构建了句内和句间两个层级的篇章分析器，在篇章主次识别任务上，其 F1 值达到了 68.43%。Feng 和 Hirst[11] 在其前期工作[12] 的基础上，使用线性链的条件随机场模型对微观篇章关系区域划分和主次做出了识别，其正确率分别达到了 85.7%和 71%。在汉语篇章树库（CDTB）上，Chu[13] 使用了上下文、词对、词和词性等特征，在主次关系识别上达到了 53.21%的正确率，识别中心在前、中心在后、多中心三类关系的 F1 值分别达到了 51.58%、53.59%、54.64%。李艳翠[14] 构建了一个自底向上的汉语微观篇章结构分析平台，其中在篇章单位主次区分的任务上，中心在前、中心在后、多中心三类识别上分别取得了 43.6%、51.5%、79.3%的 F1 值，识别的总正确率为 69%。在宏观篇章主次关系计算模型方面，还尚不完善。

3 宏观汉语篇章树库（MCDTB）

基于以上针对宏观篇章主次关系研究现状的分析，目前宏观篇章主次关系在理论、语料库建设和计算模型上还尚不完善。为此，本文构建了一个以篇章主次关系为载体的篇章结构表示体系，如图 2 所示，自上而下由全文标题、章节、段落、句子、子句等组成。其宏观结构和微观结构均是多层的，在微观篇章主次关系方面复用了李艳翠[14] 基于连接依存树的篇章结构表示体系，本文关注的重点是宏观篇章主次关系的识别模型，即段落层以上的篇章主次关系识别模型。

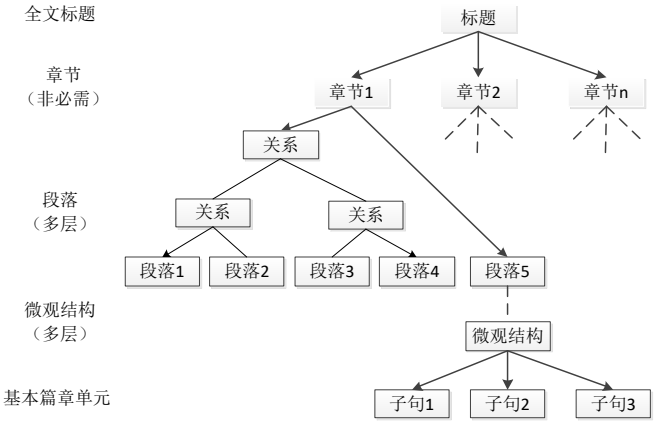

图 2 篇章结构多层树形表示

依据这个篇章结构表示体系，本文构建了宏观汉语篇章树库（MCDTB）。该语料来源于 LDC2013 年发布的 CTB8.0，选择其中最为规范的新闻报道（Newswire）作为原始语料，标注了篇章主题、篇章摘要、段落主题、篇章关系、主次关系等信息。MCDTB 更侧重在整个篇章层面，以段落为基本篇章单元，并对段落及更高层次的篇章单元间的结构、主次与关系进行相应的标注。在段落及更低的语义单元内，复用 CDTB 所标注的微观篇章结构。

表 1 宏观汉语篇章树库宏观篇章标注情况

文档总数（篇）	97
段落总数（段）	533
平均段落数（段/篇）	5.49
最大段落数（段）	13
最小段落数（段）	2
句子总数（句）	1339
平均段落长度（句/段）	2.51
段落关系总数（以二元关系计）	438

在标注方法上，本文使用采用自下而上的标注策略，在判断篇章单元的主次关系时，注重宏观上篇章单元与篇章主题的语义关联程度。在标注规则上，本文制定了一系列标准，在实施过程中，由 3 名标注人员根据标注标准对语料进行同时标注，在遇到标注不同的情况时，3 名人员经过讨论后，把一致同意的结果作为标准标注。宏观汉语篇章树库（MCDTB）目前已标注了 97 个篇章的宏观篇章结构（选取 CTB8.0 语料中前 100 篇，去掉段落数为 1 的不能形成段落间关系的 3 篇），共标注了 533 个段落之间 438 个关系（其中多元关系都转换为二元关系保存），统计数据如表 1 所示。

```
<? Xml version="1.0" encoding="UTF-8"?>
<DOC>
  <DISCOURSE DiscourseTopic="中国高新技术开发区发展迅速成果显著">
    <LEAD>中国高新技术开发区发展迅速，已形成一大批机制灵活、适应市场经济要求、技术创新能力强的高新技术企业。
</LEAD>
    <ABSTRACT>中国高新技术开发区发展迅速，已形成一大批机制灵活、适应市场经济要求、技术创新能力强的高新技
术企业，技术工贸年总收入、利税总额、年出口创汇，均比创办初增长数十倍。</ABSTRACT>
  </DISCOURSE>
<RELATION>
    <R ID="1" StructureType="逐层切分" Layer="1" RelationNumber="单个关系" RelationType="果因关系"
ParagraphPosition="1...2|3...5" Center="1" ChildList="3|2" ParentId="-1" RelationWeight="0" />
    <R ID="2" StructureType="并列切分" Layer="2" RelationNumber="单个关系" RelationType="并列关系"
ParagraphPosition="3...3|4...4|5...5" Center="3" ChildList="" ParentId="1" RelationWeight="0" />
    <R ID="3" StructureType="逐层切分" Layer="2" RelationNumber="单个关系" RelationType="解说关系"
ParagraphPosition="1...1|2...2" Center="1" ChildList="" ParentId="1" RelationWeight="0" />
</RELATION>
  <TEXT>
    <P ID="1" ParagraphTopic="中国高新技术开发区发展迅速成果显著。" ParagraphWeight="0" />
    <P ID="2" ParagraphTopic="中国高新技术开发区发展情况。" ParagraphWeight="0" />
    <P ID="3" ParagraphTopic="中国高新技术开发区企业研究开发投入情况。" ParagraphWeight="0" />
    <P ID="4" ParagraphTopic="中国高新技术开发区初步形成较为完善的支撑服务体系。" ParagraphWeight="0" />
    <P ID="5" ParagraphTopic="中国中央和地方颁布高新区管理办法。" ParagraphWeight="0" />
  </TEXT>
```

图 3 标注语料实例保存结果（chtb_0056.nw.raw）

在标注格式上，宏观汉语篇章树库（MCDTB）采用 XML 格式存储，以篇章的主题（DiscourseTopic）、短摘要（LEAD）、长摘要（ABSTRACT）、篇章关系（RELATION）、段落主题句（ParagraphTopic）为标注对象，并针对篇章关系标注了篇章关系层级（Layer）、篇章关系类型（RelationType）、篇章关系主次（Center）、篇章关系位置（ParagraphPosition）、父关系结点（ParentId）和子关系结点（ChildList）等，具体形式如图 3 所示

篇章主次关系经过二元关系转换后，具体的统计数据如表 2 所示。通过表中数据可以看出，宏观篇章主次类型 S-P 的数目十分稀少，只占到了全部数据的 4.79%。而 P-S 类型和 M-P 类型数量大致相当。考虑到本任务是识别汉语宏观篇章主次关系，根据李锦和廖开洪[15]的统计，在汉语文章中，篇章单元中重要部分在前的情况占 70%。而本文使用的新闻类篇章，因为体裁原因，主要内容在前若干段描述的比例更大，符合自然分布规律，因此本文未对数据进行的不平衡问题进行处理。

表 2 篇章主次关系统计表

篇章主次关系类型	数目	占比
P-S	239	54.57%
S-P	21	4.79%
M-P	178	40.64%
总数	438	100%

4 宏观篇章主次关系识别框架

4.1 宏观篇章主次关系计算模型

在处理宏观篇章主次关系上，本文把篇章主次关系的识别看作是一个分类问题。篇章主次关系中，多数情况都是二元主次关系，本文用一个元组来表示（[Arg1,Arg2],label），其中

Arg1和Arg2表示一个篇章主次关系的两个篇章单元，label表示两个篇章单元间的主次关系，正如图 1 中的关系 *R*3。但是也存在像 *R*2 这样的多元主次关系，本文用（[Arg1,Arg2,…,Argn],label）来表示。为了统一化表示篇章主次关系，本文把所有的多元关系都转换为二元关系。以 *R*2 为例，其元组表示为（[*c, d, e*],M-P），经过转换后，其表示形式为（[*c, d*],M-P）、（[*d, e*],M-P）。

这样最终的问题就转换为给定篇章单元 Arg1 与 Arg2，判断两个篇章单元之间的主次关系的三分类（P-S、S-P、M-P）问题。

在特征选取上，Joty[9] 、Feng[11] 、Chu[13] 等使用的为词汇、句法、文本结构等信息作为特征，而没有使用语义信息，并且上述研究都是在句内和句间进行主次关系的识别，即微观篇章主次关系。

本文的研究重点是宏观篇章主次关系，其篇章基本单元是以自然分割的文章段落，相较于微观篇章主次关系的研究，更应该注重段落之间的语义关系。考虑到词及词性等特征相对于段落的语义来说，颗粒度较小，而篇章单元与主题的相似度可以在更高层次上表现出篇章单元所涵盖的主要语义信息，因此本文将篇章单元与篇章主题的相似度作为一个重要特征，并提出了基于 word2vec[16]和基于 LDA[17]的两种主题相似度的计算方法。

4.2 基于 word2vec 的主题相似度算法

基于 word2vec 的主题相似度是计算篇章单元 Arg1 与篇章主题的语义相似度 *Score*1、篇章单元 Arg2 与篇章主题的语义相似度 *Score*2。该算法使用 word2vec 算法得到 w2vCTB 模型，再通过该模型获取目标词向量，在徐帅[18]的句子与句子之间的语义相似度计算方法的基础上，使用式（1）、（2）、（3）实现篇章单元与篇章单元的语义相似度计算，分别得出两个篇章单元与篇章主题之间的语义相似度。w2vCTB 模型使用的训练语料为 CTB8.0 前 5558 篇文章。表 3 为主题相似度获得过程中部分符号所表示的含义。

表 3 主题相似度相关符号及含义

名称	含义
w2vCTB 模型	使用 CTB8.0 前 5558 篇文章训练的 word2vec 模型。
*Score*1	Arg1 与篇章主题的语义相似度。
*Score*2	Arg2 与篇章主题的语义相似度。

在 MCDTB 语料库的宏观篇章关系中，本文把篇章标题作为篇章主题，由此计算两个篇章单元与篇章主题的语义相似度。记篇章标题为篇章单元 Arg0，则要计算的为篇章单元 Arg0 与篇章单元 Arg1 和篇章单元 Arg0 与篇章单元 Arg2 之间的语义相似度。

如式（1）所示，定义两个单词的语义相似度为余弦相似度 *Similarity*（W_i,W_j），其中 V_i、V_j 分别为单词 W_i、W_j 通过 w2vCTB 模型获得的词向量。如式（2）所示，定义篇章单元 *i* 里的第 *n* 个单词对于篇章单元 *j* 的最大映射相似度为 $MaxSim_{inj}$。如式（3）所示，定义篇章单元 *i* 和篇章单元 *j* 间的语义相似度为 *Score*。

$$Similarity(W_i,W_j)=\frac{V_i\times V_j}{|V_i|\times|V_j|} \quad (1)$$

$$MaxSim_{inj}=\max_{W_k\in C_j} Similarity(W_{in},W_k) \quad (2)$$

$$Score=\frac{\sum_{n=1}^{|C_i|} MaxSim_{inj}+\sum_{m=1}^{|C_j|} MaxSim_{jmi}}{|C_i|+|C_j|} \quad (3)$$

4.3 基于 LDA 的主题相似度算法

基于 word2vec 的主题相似度算法使用篇章标题作为篇章主题，当篇章标题不能较好的

表现出真正的篇章主题时，就会出现主题偏差现象。为了弥补这一偏差，本文提出了基于LDA 的主题相似度算法。

基于 LDA 的主题相似度是计算篇章单元 Arg1 与该篇章单元所在的篇章全文 Textall 的相似度 *LDAScore*1、篇章单元 Arg2 与该篇章单元所在的篇章全文 Textall 的相似度 *LDAScore*2。LDACTB 模型是使用 Hoffman[19]提出的 LDA 算法对 CTB8.0 中全部的新闻语料（篇章编号为 0001-0325、0400-0454、0500-0540、0600-0885、0900-0931、4000-4050）训练所得。本文使用训练好的 LDACTB 模型对篇章单元 Arg1、篇章单元 Arg2 和篇章全文 Textall 进行主题分类，并选取分类结果中概率最大的前四个主题种类作为其主题集合 $ThemeSet_1$、$ThemeSet_2$ 和 $ThemeSet_{all}$。*LDAScore*1、*LDAScore*2 的计算方法如式（4）、（5）所示。

$$LDAScore1=\frac{|ThemeSet_1 \cap ThemeSet_{all}|}{|ThemeSet_{all}|} \quad (4)$$

$$LDAScore2=\frac{|ThemeSet_2 \cap ThemeSet_{all}|}{|ThemeSet_{all}|} \quad (5)$$

4.4 特征选择

在宏观篇章主次关系分类的任务上，由于目前还未有相应的基准系统，本文使用了 Joty[9] 、Feng[11] 、Chu[13] 中使用的部分组织结构特征作为基准系统来进行比较，并在基准系统的基础上添加了基于 word2vec 和 LDA 的主题相似度作为语义特征，记基于 word2vec 的主题相似度特征为 Sim_{W2V}，基于 LDA 的主题相似度特征为 Sim_{LDA}，因此最终使用了表 4 所示的 3 组特征。

表 4 本实验使用的特征集合

组织结构特征（6 个）
篇章单元 Arg1 的开始位置和结束位置
篇章单元 Arg2 的开始位置和结束位置
篇章单元 Arg1/Arg2 各自所包含的段落数
***Sim*$_{W2V}$（2 个）**
篇章单元 Arg1 与篇章主题 Arg0 的语义相似度
篇章单元 Arg2 与篇章主题 Arg0 的语义相似度
***Sim*$_{LDA}$（2 个）**
篇章单元 Arg1 与篇章全文的 LDA 相似度。
篇章单元 Arg2 与篇章全文的 LDA 相似度。

5 实验

5.1 实验设置

本实验使用自然语言处理工具（NLTK）中的最大熵分类器（nltk.classify.maxent）[1]构建了宏观篇章主次关系识别模型，参数均使用默认选项。数据集大小为 438 条宏观篇章关系，考虑到小样本训练集的不稳定性，实验采用了十倍交叉验证，即把原数据集按照类别比例均分为 10 份，其中 1 份作为测试集，剩余 9 份作为训练集，并进行 10 次实验。

本实验使用四组不同的特征集组合进行对比验证。基准系统使用组织结构特征，第二组和第三组在基准系统的特征上分别添加了基于 word2vec 的主题相似度和基于 LDA 的主题相似度作为语义特征，第四组则在基准系统基础上，同时添加了基于 word2vec 的主题相似度特征和基于 LDA 的主题相似度特征。

5.2 实验结果

[1] http://www.nltk.org/

最终的实验结果如表 5 所示，表中的准确率（Precision），召回率(Recall)，F1 值(F1-score)分别是 3 种主次关系分类结果中标准 Precision、Recall 和 F1-score 的加权平均，正确率（Accuracy）为使用式（6）计算所得。

$$\text{正确率}=\frac{\text{主次关系样本正确分类数}}{\text{主次关系样本总数}} \qquad (6)$$

表 5 采用不同特征集的实验结果（10 次实验平均结果）

特征集	准确率	召回率	F1 值	正确率
基准系统	77.8%	79.9%	78.2%	80.01%
基准系统+Sim_{W2V}	78.4%	81.2%	79.2%	81.14%
基准系统+Sim_{LDA}	78.5%	81.5%	79.4%	81.37%
基准系统+Sim_{W2V}+Sim_{LDA}	**79.0%**	**81.9%**	**79.9%**	**81.82%**

从表 5 中可以看出，使用了组织结构、基于 word2vce 的主题相似度和基于 LDA 的主题相似度特征的第四组在准确率、召回率、F1 值和正确率上均达到最好值，相较于未添加语义特征的基准系统，分别提升了 1.2%、2.0%、1.7%和 1.81%。

第二组和第三组较基准系统都有了一定的性能提升，这证明了语义特征对于宏观篇章主次的识别具有积极作用。而融合了两种主题相似度的第四组最终取得最好性能，其原因是基于 word2vec 的主题相似度和基于 LDA 的主题相似度在不同维度上计算语义相似度，两者在语义层面形成互补，因而增强了模型识别宏观篇章主次关系的能力。

表 6 第四组的分类结果情况统计表（10 次实验平均结果）

篇章主次类型	准确率	召回率	F1 值	数量
P-S 类型	86.8%	86.3%	85.9%	23.9
S-P 类型	0%	0%	0%	2.1
M-P 类型	78.0%	85.4%	81.0%	17.8

但是，对于取得最好性能的第四组来说，不同的篇章主次类型，其表现也并不相同。如表 6 所示，各类别的情况表现出一种不平衡的分布。对于样本数量稀少的 S-P 类型，模型基本没有识别出此类别，通过对实验结果的分析后发现，一方面是因为其样本数量较少，模型没有学习到应有的特征。另一方面，S-P 类型两个篇章单元包含的段落数大致相等，因此从结构上，容易被判别为 M-P 类型。另外，S-P 类型多半为因果关系或者评价关系，对于一个篇章而言，事件的原因重要还是事件的结果重要，或者事件本身重要还是事件评价重要，通过主题相似度很难区分，在人工进行语料标注时，也存在一定的主观误差。

表 7 第四组实验结果的混淆矩阵（10 次实验平均结果）

真实值 \ 预测值	P-S 类型	S-P 类型	M-P 类型
P-S 类型	20.6	0	3.3
S-P 类型	0.7	0	1.4
M-P 类型	2.6	0	15.2

相比之下，在 P-S 类型和 M-P 类型的识别效果较为良好，通过表 7 的混淆矩阵可以看出，P-S 类型和 M-P 类型没有误分类到 S-P 类型中，除了 S-P 类型被误分类到 P-S 和 M-P 类型外，本模型的性能损失主要在于 P-S 类型与 M-P 类型之间的混淆。

6 结论与展望

实验结果证明，在宏观篇章主次关系识别的任务上，主题相似度特征能够表现出各篇章单元与篇章主题之间的密切程度，提升了宏观篇章主次关系识别的性能。本文提出的融合了基于word2vec的主题相似度和基于LDA的主题相似度的主次关系识别方法在实验中取得了最好的性能表现，其准确率、召回率、F1 值和正确率分别达到了 79.0%、81.9%、79.9%和

81.82%，相比较只含有组织结构特征的基准系统，分别提升了 1.2%、2.0%、1.7%和 1.81%。在接下来的工作中，我们将继续标注 MCDTB 语料库，完善标注规则，扩大数据集，并针对不平衡数据集出现的原因及应对策略、寻找富文本特征集等问题进行相应的探究。

参考文献

[1] 褚晓敏,朱巧明,周国栋. 自然语言处理中的篇章主次关系研究[J]. 计算机学报, 2017, 40(4): 842-860.

[2] Zou B, Zhou G, Zhu Q. Negation Focus Identification with Contextual Discourse Information[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014, 522-530.

[3] Cohan A., Goharian N. Scientific article summarization using citation-context and article's discourse structure[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015, 390-400.

[4] Liakata M., Dobnik S., Saha S., Batchelor C., Rebholz-Schuhmann D. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013, 747-757.

[5] Mann W C, Thompson S A. Relational propositions in discourse[J]. Discourse processes, 1986, 9(1): 57-90.

[6] Mann W C, Thompson S A. Rhetorical structure theory: A theory of text organization[J]. Text-Interdisciplinary Journal for the Study of Discourse, 1987, 8(3):243-281.

[7] Van Dijk T A. Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition[M]. Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates, Inc., Publishers. , 1980.

[8] Hernault H, Prendinger H, Ishizuka M. HILDA: A discourse parser using support vector machine classification[J]. Dialogue & Discourse, 2010, 1(3): 1-33.

[9] Joty S R, Carenini G, Ng R T, et al. Combining Intra-and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: Association for Computational Linguistics, 2013, 486-496.

[10] Joty S, Carenini G, Ng R T. A novel discriminative framework for sentence-level discourse analysis[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island Korea: Association for Computational Linguistics, 2012, 904-915.

[11] Feng V W, Hirst G. A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing[C]// Proceedings of the 52nd Annual Meeting of the Association for Association for Computational Linguistics. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014, 511-521.

[12] Feng V W, Hirst G. Text-level discourse parsing with rich linguistic features[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju, Republic of Korea: Association for Computational Linguistics, 2012, 60-68.

[13] Chu X, Wang Z, Zhu Q, et al. Recognizing Nuclearity between Chinese Discourse Units[C]//

Asian Language Processing (IALP). 2015 International Conference on. IEEE, 2015, 197-200.
[14] 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, 2015.
[15] 李锦,廖开洪. 汉英语篇主题与段落结构模式的比较研究[J]. 暨南学报(哲学社会科学版),2001,23(5):89-93.
[16] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.
[17] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
[18] 徐帅. 面向问答系统的复述识别技术研究与实现[D].哈尔滨:哈尔滨工业大学,2009.
[19] Hoffman M, Bach F R, Blei D M. Online learning for latent dirichlet allocation[C]// Advances in Neural Information Processing Systems. Hyatt Regency, Vancouver CANADA: Neural Information Processing Systems Foundation, Inc., 2010, 856-864.

作者联系方式：
蒋峰，江苏省苏州市干将东路 333 号苏州大学，215006，18051830985，fjiang@stu.suda.edu.cn