文章编号：1003-0077（2015）00-0000-00

基于框架的汉语篇章结构生成和篇章关系识别 *

苏娜[1]，吕国英[1]，李茹[1,2]，王智强[1]，柴清华[3]

（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；

3. 山西大学 外国语学院，山西 太原 030006）

摘要：针对汉语篇章分析的三个任务：篇章单元切割、篇章结构生成和篇章关系识别，本文提出引入框架语义进行分析研究。首先基于框架构建了汉语篇章连贯性描述体系以及相应语料库，然后抽取句首、依存句法、短语结构、目标词、框架等特征，分别训练基于最大熵的篇章单元间有无关系分类器和篇章关系分类器，最后采用贪婪算法自下向上生成篇章结构树。实验证明，框架语义可以有效切割篇章单元，并且框架特征可以有效提升篇章结构以及篇章关系的识别效果。

关键词：篇章单元；篇章结构；篇章关系；贪婪算法

中图分类号：TP391　　**文献标识码：**A

Frame-Based Discourse Structure Construction and Relation Recognition for Chinese Sentence

Na Su[1]，Guoying Lv[1]，Ru Li[1,2]，Zhiqiang Wang[1]，Qinghua Chai[3]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi, 030006, China;

2.Key laboratory of Computation Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China;

3.School of Foreign Languages, Shanxi University, Taiyuan, Shanxi, 030006, China;)

Abstract: Frame semantics was introduced to the research of Chinese discourse analysis which includes three subtasks: discourse segmentation, discourse structure construction and discourse relation recognition. First of all, the Chinese discourse coherence system and corpus was built based on frame semantics, and then two kinds of maximum entropy classifiers were applied to recognize the relation existence between discourse units and the class of discourse relation with lexical feature, dependency parser features, syntactic parser feature, target feature and frame sematic feature respectively. Finally, we used probability of the relation existence between discourse units to construct the discourse structure by greedy bottom-up method. Experimental results show that frame sematic can segment discourse units effectively and frame sematic feature can improve the performance of discourse structure construction and discourse relation recognition.

Key words: Discourse units; Discourse Structure; Discourse Relation; Greedy Bottom-up Method

1 引言

篇章分析是自然语言处理领域的一项重要任务，它[1]是指对篇章结构以及结构中篇章单元之间的语义关系进行分析。篇章由一个以上的语段或句子构成，例如，给定一个由一个句子“中国梦只有被世人理解和接受，才能加快实现进程。”构成的简单篇章，通过篇章分析后，得到如图 1 所示的篇章关系结构树。在结构树中，“中国梦只有被世人理解和接受”和“才能加快实现进程”两个篇章单元在条件关系基础上构成了一个只有一个层次的篇章结构树。该项研究对自然语言处理的许多领域起到了很大的作用，如问答系统[2]、文本连贯性[3]等。

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金项目（No.61373082）；山西省科技基础条件平台建设项目（No.2014091004-0103）；山西省回国留学人员科研资助项目（No.2013-015）；国家 863 计划项目（No.2015AA015407）；中国民航大学信息安全测评中心开放课题基金项目（No.CACC-ISECCA-201402）

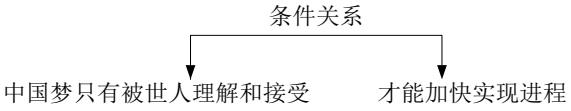

图 1 简单篇章分析示例

目前，针对篇章分析的研究主要面向英语，其中一个原因就是英文的相关理论体系和语料库比较完善。Mann 和 Thompson 提出的修辞结构理论（Rhetorical Structure Theory，RST）[1,4]认为所有好的篇章都是在篇章关系基础上形成的篇章层次化结构。基于 RST 的篇章分析器自动构建过程主要有两个子任务：（1）切割基本篇章单元（2）根据 RST 确定篇章单元之间的语义关系，生成有层次的篇章结构树。目前，已有许多研究者针对这两个任务在修辞结构理论篇章树库（Rhetorical Structure Theory - Discourse TreeBank，RST-DT）[5]上展开了研究和实验。在基本篇章单元分割任务上，Hernault[6]等人将该任务看作序列化标注问题，使用词汇、句法等平面特征训练 CRF 模型，已取得了 94%的 F 值。在篇章结构生成任务上，Wei Feng[7]等人提出使用双线性链条件随机场模型和贪婪策略进行篇章分析的方法，得到了 58.2%的正确率。

宾州篇章树库（Penn Discourse Treebank，PDTB）[8]主要标注与英语篇章连接词相关的篇章关系。基于 PDTB 的篇章分析器自动构建过程主要有三个子任务：（1）判定篇章中的连接词是否充当连接词；（2）识别存在篇章关系的两个论元（arg1，arg2）；（3）篇章关系识别，在 PDTB 中，篇章关系细分为隐式关系（Implicit）、显式关系（Explicit）、替代关系（AltLex）、实体关系（EntRel）、无关系（NoRel）5 类。篇章关系识别方面，由于显式篇章关系具有篇章连接词，易于识别，Pilter[9]等人仅仅利用连接词的统计特征已取得了 93.09%的显式篇章关系识别准确率。Ziheng Lin[10]等针对 PDTB 的第二层语义进行识别提出了短语结构树、依存句法树、上下文、词对等有效特征取得了 40.2％的隐式篇章关系识别准确率。

在汉语方面，孙静 [11]等人在自建的汉语语料库（Chinese Discourse Treebank,CNDB）上进行了相关实验。张牧宇 [12,13]等人在从 OntoNotes4.0 中随机筛选出 1096 篇文本构成的语料库上进行了相关研究与实验。涂眉 [14]等人在标有复句逻辑语义关系的清华汉语树库上，提出了基于最大熵的汉语篇章结构分析方法。但是，相对于英语篇章分析的快速发展，汉语的研究还很少，其中的主要原因是相关的理论体系与汉语篇章语料库还不够完善，且汉语在构建篇章上与英语有较大差异，使得英语的标注体系和分析方法不能完全应用到汉语上。因此，本文尝试将框架语义学与汉语篇章分析相结合，构建了相应的理论体系以及篇章框架语料库。

虽然面向篇章分析的理论以及语料库不尽相同，但从他们的实验中，可以看出句首、短语结构、依存句法等一些篇章浅层特征对篇章分析具有很大的作用。然而，篇章分析是一项艰巨的任务，仅依靠这些浅层特征还不能有效完成篇章分析任务，Ziheng Lin[10]等人曾指出识别篇章关系的难点在于歧义性、推理、上下文、世界面，篇章分析只有在分析了篇章上下文知识、理解了有联系的篇章单元的语义、对篇章单元间的语义进行合理推理等的基础上，才能分析出篇章单元之间的语义关系以及篇章的结构。Fillmore[15]的框架语义学是对世界知识和语言知识之间关系的描写，用框架对篇章进行分析，既可以在一定程度上模拟篇章的语义内容，使其具有可计算性，而且为篇章连贯提供了新的描写机制，从而有效改善篇章分析的性能。基于此，本文在框架语义基础上构建了篇章连贯性描述体系以及相应语料库，并展开了初步的句子级实验，验证了框架在汉语篇章单元切割、句子级篇章结构生成以及篇章关系识别上的作用，为进一步研究框架在篇章分析技术方面的作用奠定了基础。本文的具体组织结构如下：2、汉语篇章框架语料库介绍；3、构建篇章分析器；4、实验设置与结果分析；5、结语。

2 汉语篇章框架语料库介绍

本文利用山西大学在 Fillmore[15]提出的框架语义学理论基础上构建的汉语框架网（Chinese Framenet，简称 CFN）[16,17]，建立了方便计算机实现的篇章框架连贯性描述体系。本体系将篇章看作是由裹挟在语言符号中的框架构成的框架集合，即框架可以构成篇章单元，并且这些框架依

据语义关系自底向上组合形成一棵意义上连续的语义结构树，框架之间的语义关系通过显式或隐式的连接词语连接起来。

2.1 框架

该体系认为篇章是由裹挟在句子等表层语言符号中的框架构成的框架集合。CFN 中的框架提供了汉语词语在语言中使用的背景和动因，是人类在理解语言时，储存在人类认知经验中的图式化场景。框架语义学根据各框架对应的场景，将具有相同基本意义、支配相同类型语义角色的词语归入一个框架，比如“包含”框架下的词语有“包含”、“构成”、“涵盖”等，描述的是部分包含在整体中。篇章中裹挟在句子中的目标词（目标词是指在一个具体的句子中能够激起框架的词）激起一个与句子情境相一致的框架，句子的其它成分充当该框架的语义角色，如：

示例 1：“典型的两栖动物包括青蛙、蟾蜍、蝾螈和火蜥蜴。”进行框架语义分析后得：<tot 典型的两栖动物> <tgt=包含 包括> <par 青蛙、蟾蜍、蝾螈和火蜥蜴> 。示例 1 中的词语“包括”激活了“包含”框架，“<tot 典型的两栖动物>”、“<par 青蛙、蟾蜍、蝾螈和火蜥蜴>”是“包括”框架所支配的语义角色，其中“tot”，“par”为语义角色类型标记，分别指“整体”，与“部分”。此外，一个句子可能包含多个目标词，如“他希望专家学者持续关注、参与教育实践活动。”由 3 个目标词“希望”、“关注”、“参与”激起的框架构成。

2.2 切割篇章语义单元

针对汉语篇章由一系列句子构成，每个句子由系列小句构成的特点，本体系将一个篇章（Discourse，简称 D）中的句子经“，”、“：”等分割的语义单元定义为初级篇章单元（Primary Discourse Unit，简称 PDU），一些 PDU 没有能激起框架的目标词，即不能构成篇章的基本单元，因此将不具有框架的 PDU 与相邻具有框架的 PDU 合并在一起构成一级篇章单元(First Discourse Unit，简称 FDU)，其它含有框架的 PDU 直接向上构成 FDU；句子定义为二级篇章单元（Second Discourse Unit，简称 SDU）。这种切割方式与英语按照词汇或句法标记来划分篇章单元相比，不仅充分考虑了汉语篇章的特点，而且充分考虑了篇章单元的语义信息。

示例 2 的篇章构成如图 2 所示，例句中“()”内内容为初级篇章单元，“[]”内内容为一级篇章单元，“{}”内内容为二级篇章单元，黑体字为目标词。

示例 2：{[(今天上午)^{PDU_1}，(张乐认真**听取**发言)^{PDU_2}]^{FDU_1}，[(并与参加座谈的同志**探讨交流**)^{PDU_3}]^{FDU_2}}^{SDU_1}。{[(他**强调**)^{PDU_1}]^{FDU_1}，[(对各位专家学者**提出**的思想观点、意见建议)^{PDU_2}]^{FDU_2}，[(要认真**归纳**、**研究**、**吸收**)^{PDU_3}]^{FDU_3}}^{SDU_2}。{[(他**希望**专家学者持续**关注**、**参与**教育实践活动)^{PDU_1}]^{FDU_1}}^{SDU_3}。

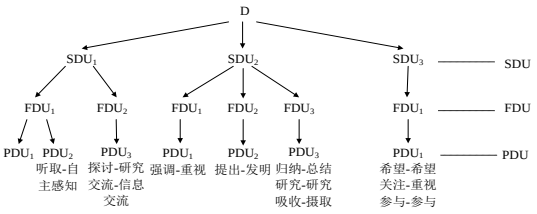

图 2 篇章语义单元构成

如图 2 所示，在 PDU 这一层级，从每个初级篇章单元中抽取出（目标词-框架），其中第一个句子的 PDU_1没有能激起框架的目标词；在 FDU 这一层级，SDU_1下的 PDU_1没有框架，与 PDU_2合并为 FDU_1，包含自主感知框架，SDU_1下的 PDU_3含有框架直接向上构成 FDU_2，包含研究、信息交流框架，篇章中其它篇章单元分析与此一致。

2.3 篇章框架结构

篇章中相邻的框架集按照篇章关系组合形成层次结构，并进一步再与相邻的层次结构组合，最终形成一棵有层次的结构树。篇章框架结构树可以用一个三元组来表示$Tree = (T_{(l,p)}, F, R)$，其中$T_{(l,p)} = (T^1{}_{(l,m)}, T^2{}_{(m+1,n)}, \dots, T^n{}_{(o,p)})$是$n(n > 0)$个篇章单元范围为$l$至$q$的篇章单元树，$T^1{}_{(l,m)}$表

示第一个篇章单元的范围为l至m，$T^2{}_{(m+1,k)}$表示第二个篇章单元的范围为$m+1$至k，$T^n{}_{(o,p)}$表示第n个篇章单元的范围为o至p，F是篇章单元范围为l至p的框架集合，$\mathrm{F}=\{f_1,f_2,\ldots,f_q\}$，$q(q>0)$，$R$表示框架集之间的关系类型，叶子节点是单个一级篇章单元中所含有的框架。

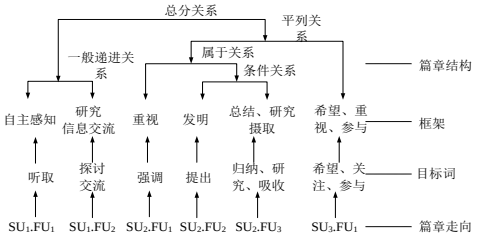

图 3 篇章框架结构图

如 2.2 的示例 2 形成的篇章结构图如图 3 所示，句子“{[(今天上午) ^{PDU_1}，(张乐认真听取发言) ^{PDU_2}] ^{FDU_1}，[(并与参加座谈的同志探讨交流) ^{PDU_3}] ^{FDU_2}} ^{SDU_1}。”具有三个基本篇章单元 PDU_1、PDU_2、PDU_3，两个一级篇章单元 FDU_1、FDU_2，其中 FDU_1和 FDU_2是一般递进关系，篇章中其它篇章单元分析与此一致。

2.4 篇章关系

本文基于黄伯荣和廖序东的《现代汉语》中关于复句以及句群之间关系分类体系[18]，建立了三层级篇章框架关系结构：第一层级根据篇章单元间意义是否平等将篇章关系划分为联合关系和偏正关系两大类别。在第二层级篇章关系中，在传统的偏正关系中加入属于关系这一类别（表 1 给出了细化至二层级的篇章关系），属于关系表示篇章的意图以及意图的所有者的所属关系。第三层级篇章关系，根据前后篇章单元的发展顺序以及逻辑关系细分为 24 类。在该篇章关系层级结构中，如果无法区分篇章单元之间的关系，将其归入承接关系的连贯关系中。

表 1 篇章关系类型

第一层	第二层	第一层	第二层
联合关系	并列关系、承接关系、递进关系、选择关系、解说关系	偏正关系	条件关系、假设关系、因果关系、目的关系、转折关系、属于关系

2.5 篇章语料库现状

鉴于目前关于汉语篇章语料库的缺乏以及标注体系的不同，我们在该理论体系下构建了一个包括 496 篇篇章的语料库，每篇文章都由人工标注了框架、篇章结构以及篇章关系。这些篇章都来自于人民日报，最小的篇章包含 1 个句子，最大的篇章包含 5 个句子，从表 2 的句子级语料库现状中，可看出总共标注了 1915 个篇章关系，其中并列关系、因果关系和属于关系所占比例较大，并列关系比例最大，达到了 21.98%；选择关系、假设关系和转折关系所占比例较小，选择关系实例数最少，只有 4 条，造成语料库这种分布状况的原因与语料体裁选取和关系本身使用频率具有较大关系。此外，三名标注人员对其中 160 篇篇章进行了同时标注，在篇章结构上取得了大于 0.9 的 kappa 值，在篇章关系上取得了大于 0.8 的 kappa 值。

表 2 句子级语料库现状

类别	并列	承接	递进	选择	解说	条件	假设	因果	目的	转折	属于	总数
数量	421	252	113	4	143	135	50	216	168	51	362	1915
比例（%）	21.98	13.16	5.90	0.36	7.47	7.05	2.61	11.28	8.77	2.66	18.9	100

3 篇章分析器

针对篇章框架语料库的篇章自动分析任务主要包括 3 个子任务：（1）根据篇章激起框架的情况，将篇章切割为一级篇章单元（FDUs）和二级篇章单元（SDUs）；（2）篇章结构生成，即生成有层次的篇章结构树。(3)篇章关系识别。为完成篇章分析的任务，本文设计了相应的篇章分析器，其具体流程如图 4 所示：

1、将进行框架分析后的篇章切割生成 FDUs 和 SDUs，以及生成篇章对应的短语结构树和依存句法树，并根据篇章单元向上组合的跨度范围与相应的短语结构树和依存语法树进行边界对齐后，分别生成训练数据集和测试数据集；

2、抽取特征训练篇章单元之间是否具有关系的最大熵分类器，对测试数据集的篇章单元对进行关系有无的预测，并利用最大熵分类模型给出的篇章单元间具有关系的概率值，采用贪婪算法生成篇章结构树；

3、抽取特征训练篇章关系分类器，对生成的篇章结构树中的篇章单元对进行关系类别预测；

4、输出标注了篇章关系的篇章框架结构树。

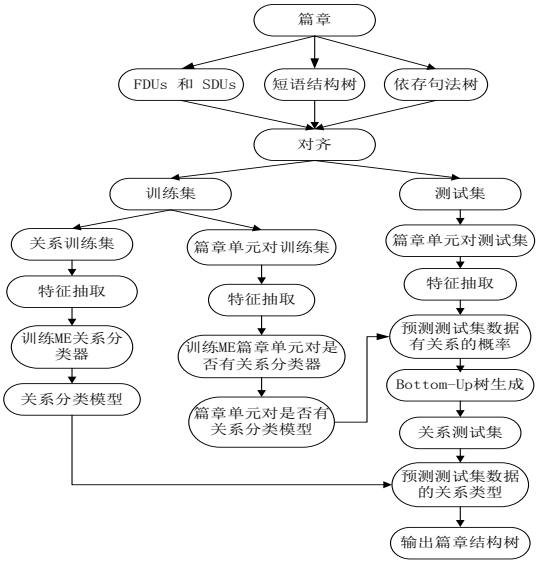

图 4 篇章分析器流程图

在训练分类器的时候，本文选用了 5 类特征：句首特征、依存句法特征、短语结构特征、目标词特征、框架特征。

3.1 特征

（1）句首

在汉语中，每个篇章单元的句首、通常起到承上启下的作用，能够起到指示篇章关系的作用。因此本文分别抽取篇章单元对的第一个篇章单元和第二个篇章单元的句首作为特征。

（2）依存句法特征

依存句法分析使用依存句法树来描述各个词语之间的语义依存关系，这种依存关系描述了篇章单元的主要信息。本文使用 Stanford Parser 对句子进行依存句法分析，然后从篇章单元向上组合的跨度范围对应的依存树中获得所有拥有被支配者的词和依存类型。图 5 显示了“张乐认真听取发言”对应的依存树，从这棵树上，收集到的依存句法特征是：听取 ← nsubj advmod dobj。每一个依存特征都表示为 3 个二元特征，来检测该特征是出现在第一个篇章单元中、第二个篇章单元中或同时出现在二者中。

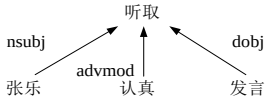

图 5 依存句法树

（3）短语结构特征

篇章单元的短语结构往往限制了篇章的结构以及篇章关系。本文使用 Stanford Parser 对每个篇章中的句子进行分析得到短语结构树，然后从这些树上提取相应篇章单元向上组合的跨度范围的短语结构特征。图 6 显示了“张乐认真听取发言”的部分短语结构树，从这棵子树上，收集到

的短语结构特征是：IP→NP VP，NP→NR，VP→ADVP VP，NR→NN，ADVP→AD 等。每一个短语结构特征都表示为 3 个二元特征，来检测该特征是出现在第一个篇章单元中、第二个篇章单元中或同时出现在二者中。

（4）目标词特征

目标词作为激起整个句子语境的词汇，在语义表达中起着很大的作用，且它们之间的关系通常反映了篇章单元间的语义关系。在 CFN 框架体系中，能承担起框架的目标词包括动词、名词和形容词。如示例 3：[第一次被严重**打击**，]$^{FDU}{}_1$ [心情相当**难过**。]$^{FDU}{}_2$。其中，FDU_1的目标词“打击”和 FDU_2的目标词“难过”代表了一种隐式的因果关系，同时也指示了 FDU_1和 FDU_2之间是因果关系。

（5）框架特征

框架能够表达文本的语义信息，选用框架作为特征不仅可以减少词语的种类，而且可以有效挖掘出框架之间的语义关系，如图 7 所示，由词语“敲打”等词语激起的框架“造成伤害”与“疼”等词语激起的框架“身体感知”是因果关系，与“惶恐”等词语激起的框架“心理刺激”同样是因果关系，除此之外，“造成伤害”框架还会与其它框架具有其它种类关系。

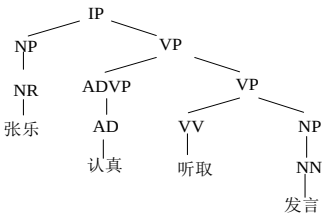

图 6 短语结构树

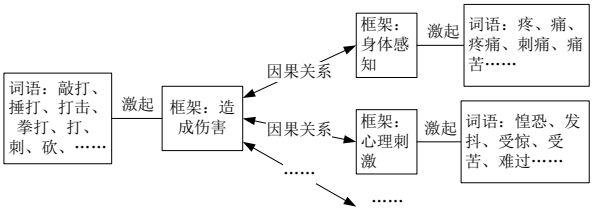

图 7 框架特征

3.2 篇章结构分析器

篇章结构分析器的任务是生成篇章相应的篇章结构树，主要包括两个子任务：（1）给篇章中任意两个相邻跨度的篇章单元是否具有关系分配概率值；（2）利用第（1）步给出的概率值采用贪婪算法生成篇章结构树。在本文中，仅考虑相邻两个篇章单元是否具有关系的现象，而不考虑跨篇章单元具有关系的现象。用$P\left(1|(T^1{}_{(l,m)},T^2{}_{(m+1,n)})\right)$来表示两个相邻篇章单元具有关系的概率，其中$T^1{}_{(l,m)}$表示第一个篇章单元的范围为$l$至$m$，$T^2{}_{(m+1,n)}$表示第二个篇章单元的范围为$m+1$至$n$。如果多个篇章单元之间具有关系，则将其分解成依次相邻两个篇章单元具有关系，如$P\left(1|(T^1{}_{(1,1)},T^2{}_{(2,2)},T^3{}_{(3,3)})\right)$，可分解为$P\left(1|(T^1{}_{(1,1)},T^2{}_{(2,2)})\right)$和$P\left(1|T^1{}_{(2,2)},T^2{}_{(3,3)})\right)$，如图 8 所示：

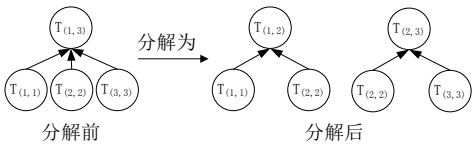

图 8 结构分解

3.2.1 篇章单元对是否有关系分类模型

在相邻篇章单元是否具有有关系的分类模型中，本文首先将篇章生成相应的篇章单元对训练集和测试集。其具体流程如下：

1、 给定篇章集合$D=\{D_1,D_2,\dots D_n\}$；

2、 将D中的元素切分为二级篇章单元集合$SU=\left\{SU_1,SU_2,\dots,SU_{(len(SU))}\right\}$；

3、 将SU 中的元素切分为基本篇章单元集合$BU=\left\{BU_1,BU_2,\dots,BU_{(len(BU))}\right\}$；

4、 判断BU中的每个元素是否包含框架，若不包含，则与相邻的元素合并，最后形成一级篇章单元集合$FU=\left\{FU_1,FU_2,\dots,FU_{(len(FU))}\right\}$；

5、 生成篇章单元对，并根据标注真实情况，为每一对篇章单元对标注是否具有关系，生成篇章结构的训练数据和测试数据。

篇章单元对是否具有关系是一个二分类问题，本文使用最大熵分类模型构建分类模型。在本实验中，用向量X表示篇章单元对$(T^1{}_{(l,m)}, T^2{}_{(m+1,n)})$，用y表示篇章单元对是否有关系，$y=1$表示具有关系，$y=0$表示没有关系，$p(y|X)$为预测X为y的概率，熵定义为：

$$H(X) = -\sum_{X,y} p(y|X) log\, p(y|X) \quad (1)$$

采用拉格朗日乘数法求解最大熵，计算公式为：

$$p(y|X) = \frac{1}{Z(X)} exp\left(\sum_{i}^{n} \lambda_i f_i(X,y)\right) (2)$$

$$Z(X) = \sum_{y} exp\left(\sum_{i}^{n} \lambda_i f_i(X,y)\right) \quad (3)$$

其中，f_i表示每个特征，n 代表特征总数，λ_i为特征的权重。

抽取框架特征、目标词特征、短语结构特征和依存句法特征来分别表示训练数据和测试数据集，用最大熵分类模型在训练集上进行训练，在测试集上进行预测后，输出详细的概率信息，将最大熵分类模型判定篇章单元之间具有关系的概率值作为两个篇章单元具有关系的概率值，即$P\big(1|(T^1{}_{(l,m)}, T^2{}_{(m+1,n)})\big) = p(1|X)$，$y=1$。

3.2.2 贪婪算法

在自下向上的结构树生成过程中，采用贪婪算法，获取每一阶段的最优值，其总体思想是将FU中的每个一级篇章单元都形成一个节点，然后依次不断比较相邻两个节点结合的概率，从中挑出概率最大的两个节点形成一个节点，直至生成的节点包含了所有的篇章单元为止。此外，篇章结构树的节点可能发生一个节点有多于两棵子树的情况，如$T_{(1,3)} = (T^1{}_{(1,1)}, T^2{}_{(2,2)}, T^3{}_{(3,3)})$，因此当两个节点生成一个新节点后，如$T^1{}_{(1,1)}$和$T^2{}_{(2,2)}$生成$T_{(1,2)} = \big(T^1{}_{(1,1)}, T^2{}_{(2,2)}\big)$，$T^2{}_{(2,2)}$与$T^3{}_{(3,3)}$又具有关系生成$T_{(2,3)} = \big(T^1{}_{(2,2)}, T^2{}_{(3,3)}\big)$，这种情况本文认定三个节点具有关系，应生成一个节点$T_{(1,3)} = (T^1{}_{(1,1)}, T^2{}_{(2,2)}, T^3{}_{(3,3)})$。在具体的构建结构树的过程中，会遇到以下 4 类情况：

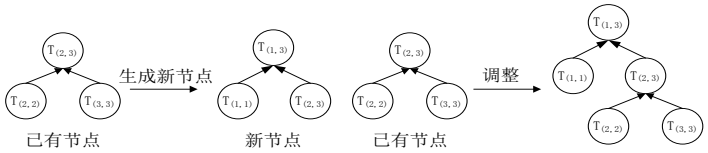

图 9 无重合

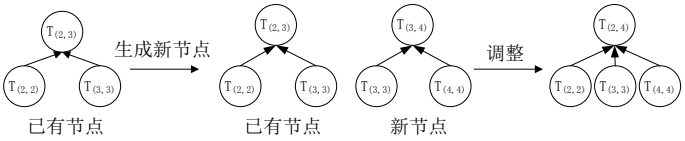

图 10 新节点的第一棵子树与已有节点的最后一棵子树相同

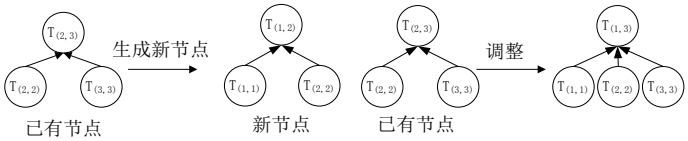

图 11 新节点的最后一棵子树与已有节点的第一棵子树相同

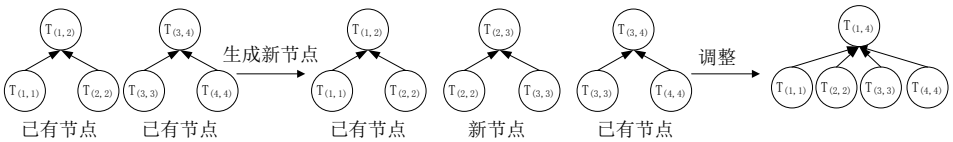

图 12 新节点的第一棵和最后一棵子树与已有节点的最后一棵子树和第一棵子树相同

第一种情况：如图 9 所示，新节点$T_{(1,3)} = \big(T^1{}_{(1,1)}, T^2{}_{(2,3)}\big)$的子树与已有节点的子树无重合，

不需要合并。

第二种情况：如图 10 所示，新生成的节点$T_{(3,4)}=\left(T^1{}_{(3,3)},T^2{}_{(4,4)}\right)$与已有节点$T_{(2,3)}=\left(T^1{}_{(2,2)},T^2{}_{(3,3)}\right)$有相同子树$T_{(3,3)}$，将其合并调整为$T_{(2,4)}=\left(T^1{}_{(2,2)},T^2{}_{(3,3)},T^3{}_{(4,4)}\right)$。

第三种情况：如图 11 所示，新生成的节点$T_{(1,2)}=\left(T^1{}_{(1,1)},T^2{}_{(2,2)}\right)$与已有节点$T_{(2,3)}=\left(T^1{}_{(2,2)},T^2{}_{(3,3)}\right)$有相同子树$T_{(2,2)}$，将其合并调整为$T_{(1,3)}=\left(T^1{}_{(1,1)},T^2{}_{(2,2)},T^3{}_{(3,3)}\right)$。

第四种情况：如图 12 所示，新生成的节点$T_{(2,3)}=\left(T^1{}_{(2,2)},T^2{}_{(3,3)}\right)$与已有节点$T_{(3,4)}=\left(T^1{}_{(3,3)},T^2{}_{(4,4)}\right)$具有相同的子树$T_{(3,3)}$，与$T_{(1,2)}=\left(T^1{}_{(1,1)},T^2{}_{(2,2)}\right)$有相同子树$T_{(2,2)}$，将其合并调整为$T_{(1,4)}=\left(T^1{}_{(1,1)},T^2{}_{(2,2)},T^3{}_{(3,3)},T^4{}_{(4,4)}\right)$。

假设一个篇章现有 4 个一级篇章单元$FU=\{FU_1,FU_2,\ldots,FU_4\}$，生成篇章结构树的具体过程如图 13 所示：

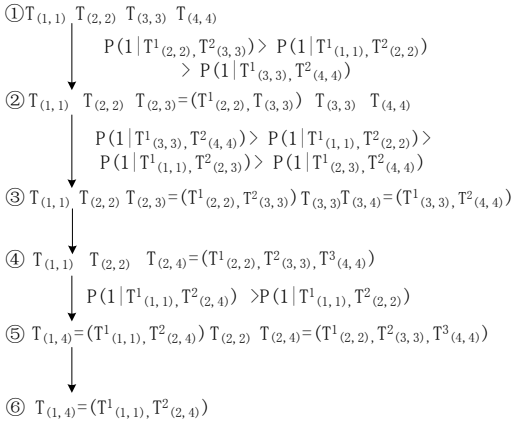

图 13 自下向上生成篇章结构树

① 将四个一级篇章单元形成叶子节点$T_{(1,1)}$、$T_{(2,2)}$、$T_{(3,3)}$、$T_{(4,4)}$；

② 根据四个节点相邻节点有关系的概率值大小，生成新节点$T_{(2,3)}=\left(T^1{}_{(2,2)},T^2{}_{(3,3)}\right)$，因为节点$T_{(2,2)}$、$T_{(3,3)}$还可能与相邻其它节点具有关系，因此保留；

③ 将②中的节点比较它们之间的概率值大小，生成新节点$T_{(3,4)}=\left(T^1{}_{(3,3)},T^2{}_{(4,4)}\right)$，因为节点$T_{(4,4)}$是最后一个篇章单元，不可能再单独与其它相邻篇章单元具有关系，因此删除；

④ ③中新生成的节点$T_{(3,4)}$造成$T_{(3,3)}$与$T_{(2,2)}$、$T_{(4,4)}$都具有关系，因此将其认定为三个篇章单元之间具有关系，将$T_{(2,3)}$、$T_{(3,4)}$合并为$T_{(2,4)}=\left(T^1{}_{(2,2)},T^2{}_{(3,3)},T^3{}_{(4,4)}\right)$，并删除节点$T_{(2,3)}$、$T_{(3,4)}$和$T_{(3,3)}$；

⑤ 将④中的节点比较它们之间的概率值大小，生成新节点$T_{(1,4)}=\left(T^1{}_{(1,1)},T^2{}_{(2,4)}\right)$，因为节点$T_{(1,1)}$是第一个篇章单元，不可能再单独与其它相邻篇章单元具有关系，因此删除；

⑥ 因为⑤中的$T_{(1,4)}$已包含所有篇章单元，因此停止比较，并删除节点$T_{(2,2)}$和$T_{(2,4)}$。从$T_{(1,4)}$开始从上向下输出这 4 个一级篇章单元生成的篇章结构树，如图 14 所示：

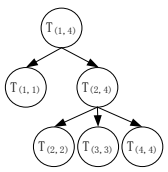

图 14 输出篇章结构树

3.3 篇章关系分类模型

针对篇章结构树中的每一对篇章单元对，依次判定它们的篇章关系类别。在判断时，选择最大熵模型作为篇章关系分类模型，使用框架特征、目标词特征、短语结构特征、依存句法特征和句首特征作为特征。当遇到多个篇章单元具有关系时，同篇章结构一样，将多个篇章单元转换为

数个相邻篇章单元对，如将$R(T^1_{(1,2)},T^2_{(3,4)},T^3_{(5,6)})$转换为$R(T^1_{(1,2)},T^2_{(3,4)})$和$R(T^1_{(3,4)},T^2_{(5,6)})$。除此之外，由于篇章关系分为显式和隐式两类，但是本语料库篇章关系数量较少，因此本文将显式篇章关系的连接词除去，转换为隐式篇章关系进行相关实验。

4 实验设置与结果分析

本文的实验分为两个部分，主要验证框架特征相较于 Ziheng Lin[10]等文献中所使用的对于篇章分析具有较大作用的短语结构树、依存句法树等特征更为有效。一是评估篇章自动分析中篇章单元对有无关系最大熵分类器和篇章关系最大熵分类器的性能，另一个是评估篇章整体性能的实验，即完全由篇章分析器自动完成篇章框架结构生成以及在结构树上标注篇章关系。在实验中用到的分类器均采用张乐的最大熵工具，使用标准正确率Acc，P，R和F值作为测试标准。

（1）篇章框架标注情况

由于 CFN 框架本身覆盖率的问题，导致篇章中的目标词不能完全标注出所属的框架，表 3 统计了篇章的框架标注情况。

表 3 框架标注情况

总标注次数	4472
标注框架次数	3679
词语总数	909
有框架的词语	679
涉及框架	193

从表 3 可以看出，总共标注了 4472 次，其中 3679 次标注了框架，所占比例为 82.27%；涉及不同词语 909 个，其中 679 个词语具有框架，所占比例为 74.70%，共涉及框架 193 个。

（2）篇章单元间有无关系识别效果

本实验采用框架特征、目标词特征、短语结构特征和依存句法特征生成篇章结构对应的特征实例集 5585 个篇章单元对，采用 5 折交叉验证进行实验，表 4 给出了每个类别特征的正确率。

表 4 基于单个特征篇章单元之间有无关系实验效果

特征	Acc./%
框架	57.48
目标词	55.81
句法	54.63
依存	44.80

表 5 基于多个特征篇章单元之间有无关系实验效果

特征	Acc./%
框架	57.48
框架+短语结构	57.53
框架+短语结构+依存句法	57.74
框架+短语结构+依存句法+目标词	57.95

通过表 4 可以看出每类特征对篇章结构分类效果的影响相继是框架特征、目标词特征、短语结构特征、依存句法特征，框架特征取得了最好的实验效果，这表明框架特征包含了更多的语义信息，更有助于识别篇章单元之间是否存在关系。

为了验证组合特征对篇章结构识别的影响，表 5 给出了特征组合对实验结果的影响。在该实验中，使用 MI 特征选择方法，选择 400 个短语结构特征、150 个依存句法特征、全部框架特征、100 个目标词特征生成篇章结构对应的特征实例集，通过表 5 可以看出组合特征的实验结果要优于单个特征，其中，框架、短语结构、依存句法和目标词特征的组合识别效果最好，这表明特征组合时，篇章结构识别效果最好。

表 6 篇章单元之间有无关系实验总效果

	P./%	R./%	F./%
有关系	40.67	26.21	31.70
无关系	63.64	76.76	69.54

表 6 给出了在所有特征组合下的篇章单元有无关系的P，R和F值。从表 6 中可以看出有关系的篇章单元对的识别效果较差，F值只有 31.70%。

（3）篇章关系实验效果

本实验采用频数大于 3 的框架特征、目标词特征、短语结构特征、句首特征和依存句法特征生成篇章关系对应的特征实例集 2110 个，采用 5 折交叉验证进行实验。为了验证各类特征在篇章关系识别上的作用，我们首先在正确标注篇章关系的数据上进行了实验，表 7 给出了各类特征的实验结果。将篇章关系中占据比例最大的并列类设置为基准系统，正确率为 22.46%。

表 7 基于单个特征篇章关系实验效果

特征	Acc./%
框架	40.69
目标词	39.45
句首	37.11
短语结构	36.39
依存句法	23.47

表 8 基于多个特征篇章关系实验效果

特征	Acc./%
框架	40.69
框架+目标词	41.48
框架+目标词+句首	46.09
框架+目标词+句首+短语结构	49.19
框架+目标词+句首+短语结构+依存句法	49.25

通过表 7 可以看出，本文选择的几组特征都是有效的，总正确率都超过了基准系统，每个特征对篇章关系分类效果的影响相继是框架特征、目标词特征、句首特征、短语结构特征和依存句法特征；框架特征的识别效果要优于目标词特征、句首特征、短语结构特征和依存树特征，达到了 40.69%，这表明标注框架对于识别篇章关系是有效的。

为了验证组合特征对实验结果的影响，表 8 给出了特征组合对实验结果的影响。通过表 8 可以看出，当所有特征组合时，实验效果最好，达到了 49.25%，比单个特征效果最好的框架提高了 8.56%，这表明组合特征时，篇章关系识别效果要明显优于单个特征。

表 9 篇章关系总效果

	并列类	承接类	递进类	选择类	解说类	条件类	假设类	因果类	目的类	转折类	属于类
P./%	46.67	39.05	10.06	-	37.13	41.00	21.08	34.56	40.00	-	85.24
R./%	63.41	37.77	4.64	-	24.08	30.75	16.39	33.22	40.47	-	92.17
F./%	53.71	38.31	6.35	-	29.02	35.02	17.90	33.72	39.78	-	88.46

表 9 分别给出了基于所有特征组合的每种篇章关系类别的P，R和F值。通过表 9 可以看出，选择类与转折类没有识别出来，假设类识别准确率较低，这是由于数据稀疏引起的，在整个语料中，选择类的实例仅有 4 个，假设类所占比重为 2.61%，转折类所占比重为 2.66%。递进类的识别效果较差，是由于递进类与并列类的特征具有较大的相似性，如若没有明显的连接词作指示，很难区分这两个类别。属于类的识别效果最好，是由于属于类别的篇章关系，多是由一些“说”、“宣布”等一些表达篇章意图的句首表达，这些词语激起了“陈述”框架，特征明显且属于类的实例数较多，对于属于类识别具有较强的针对性，因此属于类识别效果最好。并列类、承接类、解说类、条件类、因果类、目的类的识别效果相当。

（4）整体性能实验效果

为检验篇章分析器的整体性能，即完全由篇章分析器完成篇章结构生成以及在结构树上识别篇章关系，本实验首先使用贪婪策略自下向上生成篇章结构树，然后使用篇章关系分类模型对篇章结构分类模型输出的有关系篇章单元对进行关系类型预测。本实验使用 397 篇篇章作为训练集，99 篇篇章作为测试集，使用标准 Parseval[19]中的指标P，R和F值作为测试标准，实验结果如表 10。

表 10 整体实验效果

	篇章结构	篇章关系	
		标准结构	自动结构
P./%	63.77	49.39	29.95
R./%	65.35	49.39	30.02
F./%	64.55	49.39	29.99

通过表 10 可以看出，使用贪婪策略生成的篇章结构树，F值可达到 64.55%。在关系实验中，使用自动生成篇章结构的F值 29.99%比使用标准结构的F值 49.39%有所下降，这是由于自动生成的篇章结构准确率较低且篇章关系分类器的准确率也较低，以至于在下一步的自动篇章关系识别上准确率有所下降。

5 结语

本文研究了如何运用框架语义切割汉语的篇章单元以及自动分析汉语篇章结构和篇章关系。在篇章自动分析过程中，我们提出基于最大熵的分析方法，对篇章结构和篇章关系分别建模。在建模过程中使用到句首特征、依存句法特征、短语结构特征、目标词特征、框架特征，实验结果验证了框架特征可以有效提高这两个任务的准确率，为以后进一步的工作奠定了基础。但是由于本文的框架覆盖不全，造成实验效果并未达到最优，因此在以后的工作中，我们将进一步进行框架的构建工作，同时有效使用框架语义资源在汉语篇章分析方面的研究，如框架的语义角色、框架关系等，并扩大篇章单元的研究范围。

参考文献

[1] Mann W C, Thompson S A. Rhetorical structure theory: A framework for the analysis of texts.[J]. Iprapapers in Pragmatics, 1987,1: 79-105.

[2] Prasad R, Joshi A. A discourse-based approach to generating why-questions from texts[C]//Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge, Arlington, VA. 2008.

[3] 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-Volume 1. Association for Computational Linguistics, 2011: 997-1006.

[4] Mann W C, Thompson S A. Rhetorical structure theory: Toward a functional theory of text organization [J]. Text, 1988,8(3): 243-281.

[5] L. Carlson, D. Marcu. Building a discourse-tagged corpus in the framework of rhetorical structure theory [C]//Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, 2001.

[6] Hernault H, Bollegala D, Ishizuka M. A sequential model for discourse segmentation [M]//Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2010.

[7] Vanessa Wei Feng, Graeme Hirst. A linear-time bottom-up discourse parser with constraints and post-editing.[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA, June 2014:511–521.

[8] PDTB Research Group. The penn discourse treebank 2.0 annotation manual [R]. Philadelphia: University of Pennsylvania, 2008.

[9] Pitler E, Raghupathy M, Mehta H, et al. Easily identifiable discourse relations[C]// International Conference on Computational Linguistics. 2008:87-90.

[10] Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the penn discourse treebank [C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Morristown: Association for Computational Linguistics, 2009: 343-351.

[11] 孙静,李艳翠,周围栋,冯文贺.汉语隐式篇章关系识别[J].北京大学学报,2013,11.

[12] 张牧宇,宋原,秦兵,刘挺.中文篇章级句间语义关系识别[J].中文信息学报. 2014,28(2):28-36.

[13] 姬建辉,张牧宇,秦兵,刘挺.中文篇章级句间关系自动分析[J].江西师范大学学报（自然科学报）. 2015,3.

[14] 涂眉,周玉,宗成庆.基于最大熵的汉语篇章结构自动分析方法[J].北京大学学报(自然科学版)，2014,1.

[15] Fillmore, Charles J. Frame semantics [A]. In Linguistics in the Morning Calm, the Linguistic Society of Korea, Seoul: Hanshin. 1982:111-137.

[16] 李茹．汉语句子框架语义结构分析技术研究[D].山西大学博士学位论文. 2012.

[17] 郝晓燕,刘伟,李茹等.汉语框架语义知识库及软件描述体系[J].中文信息学报, 2007, 21(5): 96-100.

[18] 黄伯荣,廖序东.现代汉语[M].北京：高等教育出版社.2011.

[19] Abney S, Flickinger D, Gdaniec C, et al. Procedure for quantitatively comparing the syntactic coverage of English grammars[J]. Proceedings of the Workshop on Speech & Natural Language, 1991.

苏娜（1989-），女，硕士研究生，主要研究领域为中文信息处理。
Email:cindysunas@163.com；

吕国英（1964-），女，硕士，副教授，硕士生导师，主要研究领域为自然语言处理；
Email:english@sxu.edu.cn；

李茹（1965—），女，博士，教授，博士生导师，主要研究领域为自然语言处理。
Email:liru@sxu.edu.cn。

通讯作者：吕国英，Email: english@sxu.edu.cn