基于汉语框架语义网的篇章关系识别

李国臣[1,2]，张雅星[1]，李茹[1,3,4]

（1.山西大学 计算机与信息技术学院，山西 太原 030006；
2.太原工业学院 计算机工程系，山西 太原 030006；
3. 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 030008；
4. 山西省大数据挖掘与智能技术协同创新中心，山西 太原 030006）

摘 要：篇章关系识别是篇章分析中一项具有挑战性的子任务。传统的篇章关系分析主要是用篇章的局部特征对篇章关系进行分析，但是局部特征无法直接诠释篇章单元的外部语义关系，因此本文基于汉语框架语义网识别篇章关系，在框架语义层面对篇章单元进行分析。本文主要利用汉语框架语义网中的目标词，对篇章单元进行分析，从而识别出篇章关系。本文的实验结果表明，核心目标词更能完整表达篇章单元的核心语义，对篇章关系的识别有较好的效果。

关键词：篇章关系；汉语框架语义网；篇章单元；核心目标词

中图分类号：TP391　　**文献标识码：**A

Discourse Relation Recognition Based on Chinese FrameNet

LI Guochen[1,2]，ZHANG Yaxing[1]，LI Ru[1,3,4]

(1.School of Computer & Information Technology,Shanxi University,Taiyuan,Shanxi 030006,China;
2.Department of Computer Engineering,taiyuan Insititute of Technology,Taiyuan,Shanxi 030008,china;
3.Key Laboratory of Computation Intelligence and Chinese Information Processing of Ministry of Education,Shanxi University,Taiyuan,Shanxi 030006,China;
4.Collaborative Innovation Center of Big Data Mining and Intelligent Technology in Shanxi,Taiyuan,Shanxi 030006,China)

Abstract: Discourse Relation Recognition is a challenging sub-task in discourse analysis. The traditional discourse relation analysis aims to use the local feature of discourses to analyze the discourse relation. As the local feature cannot directly explain the external semantic relation of the discourse unites, we recognize the discourse relation based on chinese framenet and analyze it at the level of frame semantics. In this paper, we can recognize the discourse relation by analyzing the discourse unites with the targets in chinese framenet. Some experiments are made and the results show that the core target can perfectly express the core semantics of discourse unites and improve the performance of discourse relation recognition.

Key word: Discourse Relation；Chinese FramenNet；Discourse units；Core target

1 引言

篇章关系识别是篇章分析中重要的子任务，它研究的是篇章中两个篇章单元的关系。例

收稿日期：　　**定稿日期：**

基金项目：国家自然科学基金项目（No.61373082）；国家 863 计划项目（No.2015AA015407）；国家自然科学基金重点项目（No.61432011,No.U1435212）

作者简介：李国臣（1963—），男，教授，硕士生导师，主要研究方向为中文信息处理；张雅星（1992—），女，硕士研究生，主要研究方向为中文信息处理；李茹（1963—），女，教授，博士生导师，主要研究方向为中文信息处理。

如，本文给出一个简单篇章：“令人欣喜的是，现在媒体对会议进行了相当广泛的评论和报道。”通过对该篇章中的两个篇章单元进行篇章关系识别，可以得到前置篇章单元“令人欣喜的是”与后置篇章单元“现在媒体对会议进行了相当广泛的评论和报道”的篇章关系为解说关系。

目前，篇章关系的分析主要是面向英文，其中最主要的原因是英文的篇章分析理论体系比较完善。英文的篇章分析理论体系主要有修辞结构理论（Rhetorical Structure Theory，RST）和宾州篇章树库（Penn Discourse Treebank，PDTB）。

修辞结构理论[1]是由美国学者 William c. Mann 和 Sandra A. Thompson 等首创于 1988 年，是一套关于自然语篇结构描写的理论体系。基于 RST 的篇章关系识别主要有两个子任务：（1）基本篇章单元的生成；（2）根据 RST 对篇章单元之间的篇章关系进行分析。根据话语效果的位置，RST 将篇章中的修辞关系分为两个大类：并列型的“多级核心(multinuclear)关系”和主从型的“核心(nuclear)/辅助(satellite)关系”。其中并列型关系分为对比、结合、列举、多级核心重述和序列，主从型关系分为“表述”和“主题”关系。目前，已有许多学者在修辞结构理论篇章树库（Rhetorical Structure Theory-Discourse TreeBank，RST-DT）[2]上展开了研究和实验。Marcu[3]提出了一种无监督的方法来识别篇章关系，该方法从训练语料中抽取词对信息作为基本特征训练贝叶斯分类模型，其中某些句间关系分类模型取得了 93%的准确率。

宾州篇章树库[4]是主要标注与篇章连接词相关的篇章关系。宾州篇章树库根据两个篇章单元之间是否存在连接词，将篇章关系分为显示篇章关系和隐式篇章关系。其中隐式篇章关系又分为：替代词汇化（AItLex）、基于实体一致性关系（eNTRel）、没有关系（NoRel）。宾州篇章树库还另外对所有的篇章关系定义了一个三层的语义结构：最上层是种类，第 2 层是类型，最下层是子类型。其中，第 1 层包括 4 种最常见的语义：扩展（Expansion）、时序偶然（Contingency）、对比（Comparison）和时序（Temporal），第 2 层包括 16 类语义，第 3 层包括 23 类语义。在篇章关系识别方面，Pilter[5]等人在连接词识别的基础上使用朴素贝叶斯方法依据连接词和句法信息特征对第一层显式关系进行识别，其准确率达到了 94.15%。Lan[6]等人在交互结构优化多任务学习框架下，抽取论元的动词、极性等基本语言学特征训练基于现实语境的隐式论元对数据的主分类器和基于人造伪隐式论元对数据的辅分类器，提升隐式关系推理性能至 42.30%。

在汉语方面，孙静[7]等人在自建的汉语篇章结构语料库（Chinese Discourse Treebank，CDTB）上进行了隐式篇章关系的识别。张牧宇[8-11]等人在哈工大中文篇章关系树库（HIT-CDTB）上进行了篇章分析的相关研究。目前篇章关系分析方法主要采用短语结构、依从句法、词共现等一些篇章的浅层特征进行分析，虽然这些特征对篇章关系分析具有很大的作用，但是篇章关系识别是一项有挑战性的任务，仅依靠这些浅层特征不能有效的完成篇章关系识别任务。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，才能更好的分析出篇章单元之间的语义关系。因此，本文在苏娜[12]基于汉语框架语义所构建的理论体系上进行篇章关系的识别。在该理论体系中，篇章是由与该篇章内容相关的框架集组合而成。具体描述为：较小的框架集描述的场景按照篇章关系组合形成更大的场景，并进一步再与相邻的框架集所描述的场景组合，最终形成一棵具有层次的篇章框架语义结构树，描述一个完整的最大的语义场景。根据该理论体系，每个篇章单元的场景可以由框架集进行描述，因此，每个篇章单元都可以由相应的框架集代替。本文找出可以代替要分析的篇章单元的场景的框架集，用该框架集中的核心框架来代替该语义场景，因此将分析两个篇章单元间的关系改为分析两个框架的关系。而且在本文所用的方法中，用框架语义识别篇章关系，可以有效改善篇章关系识别性能。本文在第二节简单介绍了汉语框架语义网；第三节具体介绍了篇章关系识别的步骤；第四节描述了实验设置并对实验结果进行分析；第五节总结

全文并展望未来的研究工作。

2 汉语框架语义网介绍

汉语框架语义网（Chinese FrameNet，CFN）[13,14]是山西大学在 Fillmore 提出的框架语义学基础上所构建的，以加州大学伯克利分校的 FrameNet 为参照，以汉语真实语料为依据的供计算机使用的汉语词汇语义知识库。该知识库包括框架库，句子库，词元库三部分。

框架库是以框架为单位，对词语进行分类描述。框架是与一些激活性语境相一致的结构化范畴系统，它是存储在人类认知经验中的图示化情境，是理解词语的背景和动因，场景内容可以是一个动作，一个活动事件，一个实体或者一个抽象体的状态。框架承担词包括动词，形容词，名词，成语以及一些约定俗语，它们是能够激起汉语框架语义网某个框架所对应的语义场景的词语，是标注工作的着眼点，称为词元。一般情况下，一个框架包括多个词元。在实际例句中出现的可以激起框架语义场景的词元是目标词。如：

例 1：篇章单元“这位负责人表示这些年各地高度重视保障工资支付工作。”中的目标词有“表示”、“重视”。“表示”与“重视”可以激起的框架分别为“陈述”、“重视”，也即“表示”为框架“陈述”的词元，“重视”为框架“重视”的词元。以“表示”为例对该篇章单元进行分析后可得：<spkr 这位负责人> <tgt=陈述 表示> <msg 这些年各地高度重视保障工资支付工作>。“<spkr 这位负责人>”和“<msg 这些年各地高度重视保障工资支付工作>”是“陈述”框架所支配的语义角色，其中“spkr”和“msg”为语义角色的类型标记，分别指“说话者”和“信息”。

3 基于汉语框架语义网的篇章关系识别

本文基于汉语框架语义网识别篇章关系，通过使用篇章单元对的框架集合，对篇章单元对的框架对进行抽取，得到框架对关系表，将待测篇章单元对的核心目标词对对应的框架对与框架对关系表进行对照，得到待测篇章单元对的篇章关系。篇章关系识别的具体流程如图 1 所示：

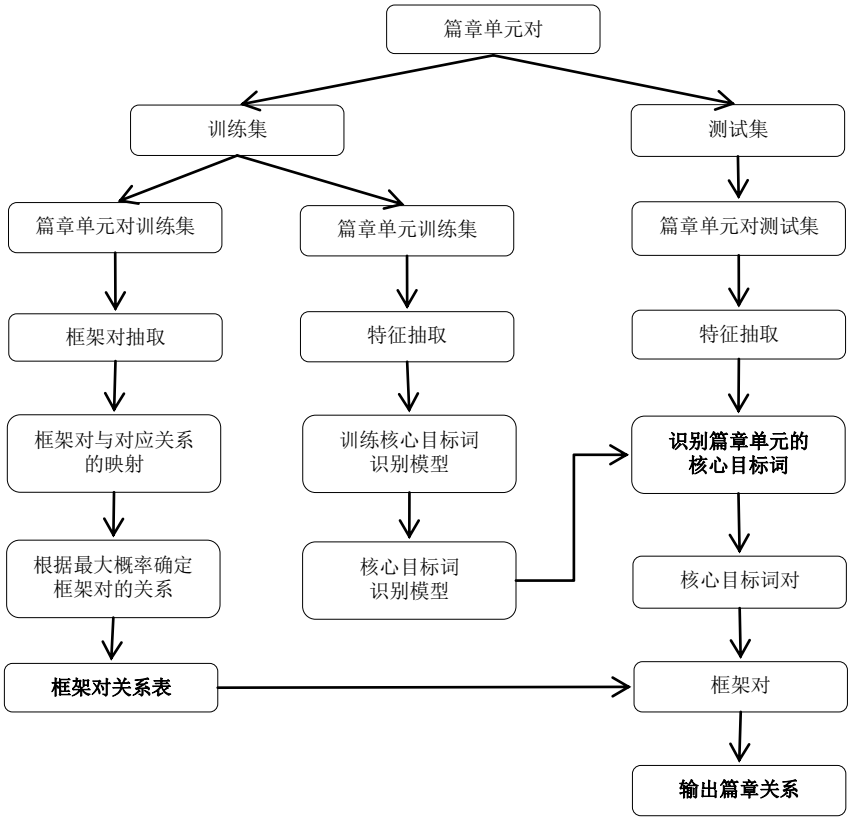

图 1 篇章关系识别流程图

本文对篇章关系的识别主要包括以下三个步骤：

（1）将已标注语料分为训练数据集和测试数据集，对训练数据集进行框架对的抽取，得到框架对与对应关系的映射，计算每个框架对的最大概率关系，生成框架对关系表；

（2）抽取特征训练核心目标词识别的最大熵模型，对测试数据集的篇章单元对进行核心目标词的识别，生成核心目标词对；

（3）将测试集的核心目标词对对应的框架对与第一步生成的框架对关系表进行对照，得到测试集对应的篇章关系。

3.1 框架对关系表生成

3.1.1 框架对抽取

对所标注语料进行框架对的抽取。抽取框架对的具体步骤为：

（1）抽取前置篇章单元的所有框架，获得框架集合 FrameSet1，FrameSet1 包含 m 个框架{$Frame_{11}$，$Frame_{12}$，....，$Frame_{1m}$}；同理，抽取后置篇章单元的所有框架，获得框架集合 FrameSet2，FrameSet2 包含 n 个框架{$Frame_{21}$，$Frame_{22}$，....，$Frame_{2n}$}；

（2）对 FrameSet1 和 FrameSet2 中的所有框架进行两两配对，形成所有可能的框架对 $\{Frame_{1i}，Frame_{2j}\}_{i=1\ldots m，j=1\ldots n}$；

（3）该篇章单元对形成的所有的框架对都对应于该篇章单元对已经标注的篇章关系；

（4）对所有的篇章单元对进行上面三个步骤，得到所有训练集形成的框架对与对应关系的映射。

以例 2 为例，将抽取框架对的步骤进行详细说明。

例 2：篇章单元对：在新的历史时期中国梦的本质是国家富强、民族振兴、人民幸福，我们的奋斗目标是到 2020 年全面实现小康社会。

前置篇章单元：在新的历史时期中国梦的本质**是**国家富强、民族振兴、人民幸福

后置篇章单元：我们的奋斗目标**是**到 2020 年全面**实现**小康社会

篇章关系：并列关系

在例 2 中，前置篇章单元和后置篇章单元包含的目标词和对应框架如表 1 所示：

表 1 篇章单元对的目标词与框架

	前置篇章单元	后置篇章单元	
目标词	是	是	实现
框架	等同	等同	实现

从表 1，可以得到前置篇章单元的框架集合 FrameSet1 为{等同}，后置篇章单元的框架集合 FrameSet2 为{等同，实现}，则对 FrameSet1 和 FrameSet2 中的框架两两配对形成的框架对为{等同，等同}、{等同，实现}。根据该篇章单元对的篇章关系为并列关系，则这两对框架对的对应关系为并列关系。对所有的篇章单元对进行如例 2 所示的步骤，得到所有训练集形成的框架对与对应关系的映射。

3.1.2 框架对的最大概率关系

将得到的所有框架对以及每个框架对在不同篇章单元对中的相应关系进行不重复合并，得到框架对与篇章关系的关系映射表 Fmap。

借助关系映射表 Fmap，本文对每种框架对最可能对应的关系进行计算。将篇章关系的十一种关系进行编号 i，$i \in \{1,2,...,11\}$。特定框架对{$Frame_{1i}$，$Frame_{2j}$}$_{i=1\ldots m,\ j=1\ldots n}$在关系映射表 Fmap 中对应这 11 种关系出现的频次分别为 r_i，$i \in \{1,2,...,11\}$，在关系映射表中出现的总数为 n。本文用 r_i 除以 n 计算特定框架对{$Frame_{1i}$，$Frame_{2j}$}$_{i=1\ldots m,\ j=1\ldots n}$在每种关系上的分布概率，其分布概率最大的数值对应的关系 r 为该框架对的篇章关系。公式如下所示：

$$i = \arg\max_{j=1}^{11}\{\frac{r_i}{n}\} \quad (1)$$

如例 1 中的框架对{等同，等同}，在关系映射表中对应递进关系出现 1 次，对应解说关系关系出现 1 次，对应因果关系出现 4 次，对应并列关系出现 9 次，其余关系类都没有出现，则出现总次数为 15 次。分别用 1，1，4，9 除以 15，可以得到概率最大的为出现次数为 9 次的并列关系，则框架对{等同，等同}对应的篇章关系为并列关系。

本文对关系映射表 Fmap 中的每种框架都进行上述计算，得到框架对关系表 FRmap。

获得 FRmap 的算法如下：

算法 1：获取框架对关系表 FRmap 算法

输入：篇章单元对集合 D={D_1，D_2，...，D_n}，每个篇章单元对 D_i 的前置篇章单元 D_{i1} 和后置篇章单元 D_{i2} 的篇章关系 R_i

输出：框架对关系表 FRmap

1. FOR D_i IN D
2. FOR D_{ij} IN D_i // $j \in \{1,2\}$
3. 获得 D_{ij} 的框架集合 FrameSetj={$Frame_{j1}$，$Frame_{j2}$，....，$Frame_{jm}$}
4. END FOR
5. FOR $Frame_{1x}$ IN FrameSet1
6. FOR $Frame_{2y}$ IN FrameSet2
7. $Frame_{1x}$ 与 $Frame_{2y}$ 配对，并将{$Frame_{1x}$，$Frame_{2y}$，R_i}放入表 Fmap
8. END FOR
9. END FOR // 得到篇章单元对 D_i 前置篇章单元的所有框架和后置篇章单元的所有框架的两两配对
10. END FOR
11. FOR $Fmap_i$ IN Fmap
12. IF ! $Fmap \in FRmap$ //只进行框架对的对照

13. 根据公式（1）计算框架对 $Fmap_i$ 的篇章关系，并将该框架对和对应篇章关系放入表 FRmap

14. END IF

15. END FOR

Return FRmap。

3.2 核心目标词识别

识别核心目标词着眼点是篇章单元中的一个词，识别该词是否是核心目标词，因此本文将这项任务看做分类问题来解决，使用最大熵模型构建分类模型。

在本实验中，用向量 X 表示篇章单元，用 y 表示候选目标词是否是核心目标词，p（y|X）为预测 X 为 y 的概率，熵定义为：

$$H(X)=-\sum_{X,y}p(y\mid X)\log p(y\mid X) \quad (2)$$

采用拉格朗日乘数法求解最大熵，计算公式为：

$$p(y\mid X)=\frac{1}{Z(X)}\exp(\sum_{i}^{n}\lambda_i f_i(X,y)) \quad (3)$$

$$Z(X)=\sum_{y}\exp(\sum_{i}^{n}\lambda_i f_i(X,y)) \quad (4)$$

其中，f_i 表示每个特征，n 表示特征总数，λ_i 为特征的权重。

抽取词形、词性、当前词前一个词的词性、当前词后一个词的词性、依从关系来分别表示训练集数据和测试集数据，用最大熵分类模型在训练集上进行训练，在测试集上进行识别，得到篇章单元的核心目标词。

3.3 篇章关系识别

将测试集中的篇章单元对进行核心目标词的识别，得到每个篇章单元的核心目标词，从而可以得到篇章单元对的核心目标词对，得到所对应的框架对。

将篇章单元对的核心目标词对对应的框架对与 FRmap 进行对照，得到该框架对对应的篇章关系。该篇章关系就是待测篇章单元对的关系。

以例 3 为例，对篇章关系的识别步骤进行说明。

例 3：篇章单元对：仅 2012 年全国共发生 0 到 12 岁儿童伤亡交通事故 11117 起，造成 12153 名儿童伤亡。

前置篇章单元：仅 2012 年全国共**发生** 0 到 12 岁儿童伤亡交通事故 11117 起

后置篇章单元：造成 12153 名儿童伤亡

如例 3 中，前置篇章单元的核心目标词是发生，所属框架为事件，后置篇章单元的核心目标词是造成，所属框架是因果。因此可以得到该待测篇章单元对的核心目标词对对应的框架对为{事件，因果}，与框架对关系表 FRmap 对照，可以得到{事件，因果}的篇章关系为承接关系，所以该篇章单元对的篇章关系为承接关系。

本文基于框架语义的篇章关系识别算法如下：

算法 2：篇章关系识别算法

输入：待测篇章单元对 D，框架对关系表 FRmap

输出：待测篇章单元对的篇章关系

1. FOR Di IN D // $i\in\{1,2\}$

2. 将 Di 经过核心目标词识别模型，识别出核心目标词 Wi

3. END FOR

4. 将核心目标词对{M1，M2}在 FRmap 中查找对应的篇章关系 R

Return R。

4 实验设置与结果分析

4.1 实验语料

4.1.1 篇章关系

本文所采用的篇章关系[12]是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而建立的。该篇章关系结构分为三层。第一层根据篇章单元之间意义是否平等分为联合关系和偏正关系。第二层中，联合关系可分为并列关系、承接关系、递进关系、选择关系、解说关系。偏正关系可分为条件关系、假设关系、因果关系、目的关系、转折关系、属于关系。该体系在传统的偏正关系中加入属于关系这一类别，属于关系表示篇章的意图以及意图的所有者的所属关系。第三层根据前后篇章单元的功能分为 24 类。在该篇章关系中，如果无法区分篇章单元之间的关系，就将其归为承接关系中的连贯关系。篇章关系如表 2 所示。

表 2 篇章关系

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

4.1.2 篇章语料库

本文研究的是相邻的两个篇章单元之间的关系，并且本文的实验方法是基于汉语框架语义网的。因此所用语料必须具有下列特点：

（1）具有前置篇章单元和后置篇章单元；

（2）前置篇章单元和后置篇章单元必须且至少包含一个可以激起框架的目标词。

本文对所获得的语料都进行了人工标注，对每对篇章单元对都标注了框架与篇章关系。这些语料主要来源于新闻语料和语料库在线。语料中各个篇章关系的分布概率如表 3 所示。

表 3 篇章语料库

关系	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	495	384	225	5	265	245	90	354	189	287	435	2974
比例（%）	16.64	12.91	7.56	0.17	8.91	8.23	3.03	11.90	6.36	9.65	14.63	100

在训练识别核心目标词模型时，本文使用哈尔滨工业大学信息检索研究中心的语言处理集成平台 LTP[15]对语料进行预处理。实验语料的统计结果如表 4 所示。

表 4 标注语料

	例句数	词元数	框架数
语料集	5548	6350	179

4.2 评价标准

本文使用准确率 Acc（Accuracy）、精确率 P（Precision）、召回率 R（Recall）和 F 值作为篇章关系识别性能的度量指标。假设 $i \in \{1,2,...,11\}$，分别对应十一种篇章关系中的一种，R_i 为实验中预测出关系为 i 的个数，C_i 为实验中预测正确的关系为 i 的个数，A_i 为测试集中关系为 i 的个数，则：

（1）计算十一种关系总的性能时，本文将准确率、精确率、召回率和 F 值表示如下：

$$Accuracy=\frac{\sum_{i=1}^{11}C_i}{\sum_{i=1}^{11}A_i} \qquad P=\frac{1}{11}\sum_{i=1}^{11}\frac{C_i}{R_i} \qquad R=\frac{1}{11}\sum_{i=1}^{11}\frac{C_i}{A_i} \qquad F=\frac{2PR}{P+R}$$

（2）分别计算每种关系的性能时，本文将准确率、精确率、召回率和 F 值表示如下：

$$Accuracy=\frac{\sum_{i=1}^{11}A_{i-(}A_i+R_i-C_{i)+}C_i}{\sum_{i=1}^{11}A_i} \qquad P_i=\frac{C_i}{R_i} \qquad R_i=\frac{C_i}{A_i} \qquad F=\frac{2PR}{P+R}$$

4.3 实验结果与分析

4.3.1 框架对关系表 FRmap 的生成

本文选用了 2774 篇篇章单元对作为训练集生成框架对关系表 FRmap，200 篇篇章单元对作为测试集。

生成的框架对关系表 FRmap 共有 2216 对不同框架对，其中十一种篇章关系的分布概率如表 5 所示。

表 5 FRmap

关系	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	362	355	103	2	451	97	37	267	189	109	244	2216
比例（%）	16.33	16.02	4.65	0.09	20.35	4.38	1.67	12.05	8.53	4.92	11.01	100

4.3.2 核心目标词的识别

本文对要测试的 200 篇篇章单元对即 400 个篇章单元经过预处理，然后用生成的核心目标词识别模型进行识别。识别结果如表 6 所示。

表 6 核心目标词识别结果

	Acc/%	P/%	R/%	F/%
核心目标词	90.87	94.99	95.12	95.05

经过分析，识别核心目标词正确率不高的原因是：训练语料无法包含所有的目标词，存在未登录词，使得核心目标词的识别存在困难。如：篇章单元对：对各位专家学者提出的思想观点、意见建议，要认真归纳、研究、吸收。识别后置篇章单元“要认真归纳、研究、吸收”的核心目标词时，经过核心目标词识别模型的识别，目标词“归纳”、“研究”、“吸收”为核心目标词的概率相同，无法准确判断核心目标词。

4.3.3 篇章关系的识别

（1）按照本文所说实验步骤进行，所得到的最终结果如表 7 所示：

表 7 篇章关系识别结果

篇章关系	Acc/%	P/%	R/%	F/%
并列关系	79.75	25.92	35.00	29.78
承接关系	84.05	18.75	18.75	18.75
递进关系	88.34	30.00	20.00	24.00
选择关系	-	-	-	-
解说关系	88.96	33.33	50.00	40.00
条件关系	91.41	40.00	33.33	36.36
假设关系	95.09	25.00	16.67	20.00
因果关系	85.27	30.77	21.05	25.00

目的关系	96.93	10.00	50.00	16.67
转折关系	94.48	33.33	28.57	30.77
属于关系	82.82	72.22	75.00	73.58
所有关系	43.55	29.03	31.67	31.71

表 7 给出了每种篇章关系类别的 Accuracy、P、R 和 F 值。通过表 7 可以看出，选择类没有识别出来，目的类和假设类的识别率较低，这是由于数据稀疏引起的，在所有语料中，选择类仅仅有 5 例，目的类所占比例为 6.36%，假设类所占比例为 3.03%。承接类和递进类的识别效率低，则是由于承接类和递进类的语义比较相近，因此比较难以区分这两个类别。属于类识别效果最好，这是由于识别类的篇章单元多由“说”、“称”、“强调”等可以激起“陈述”框架词语进行引导，而且属于类的实例也比较多，因此属于类效果最好。

（2）在测试集中，将每个篇章单元对中的框架都进行两两配对，生成框架对的步骤与生成框架对关系表 FRmap 的步骤一样。将生成的每一对框架对都和 FRmap 进行对照，得到框架对对应的篇章关系，将该篇章单元对的所有框架对对应的篇章关系进行统计，篇章关系相同的进行相加，最后出现的最多的关系为该篇章单元对的关系。所得实验结果如表 8 所示：

表 8 篇章关系识别结果

篇章关系	Acc/%	P/%	R/%	F/%
并列关系	73.27	15.38	19.35	17.14
承接关系	81.10	8.70	9.09	8.89
递进关系	89.86	40.00	20.00	26.67
选择关系	-	-	-	-
解说关系	86.18	13.64	21.43	16.67
条件关系	89.86	25.00	18.75	21.43
假设关系	-	-	-	-
因果关系	80.18	27.27	22.27	24.52
目的关系	87.10	33.33	38.46	35.71
转折关系	92.63	11.11	11.11	11.11
属于关系	81.11	65.57	66.67	66.12
所有关系	32.25	21.82	20.65	21.21

表 8 中，选择类和假设类都没有识别出来，这是由于数据稀疏引起的，在整个语料中，选择类仅仅有 5 例，假设类所占比例为 3.03%。并且通过与表 7 对比，表 7 只有选择类没有识别出来，可以看出，该方法更加依赖于语料规模的大小。下面图 2 对两个实验的精确率进行对比。

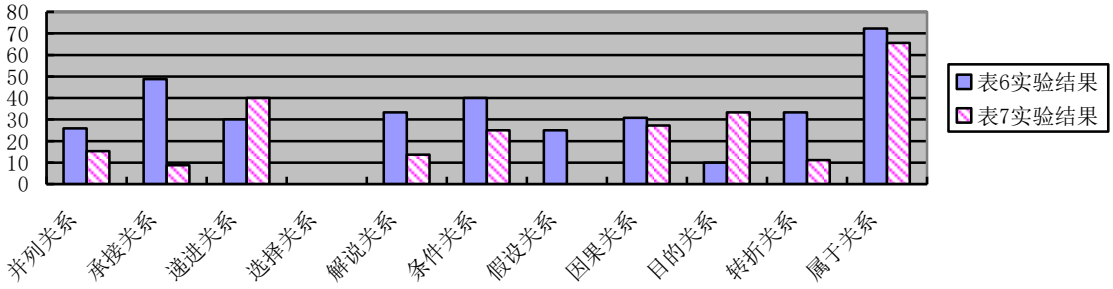

图 2 实验结果对比

通过图 2 可以看出，表 8 的篇章关系识别结果只有递进类和目的类比表 7 好，因此可以看出识别篇章单元的核心目标词可以提高识别篇章关系的准确率。这是由于表 8 所示的实验采用的简单配对的方法，触发核心框架的概率小，所形成的框架对无法较好的表达篇章单元

的核心语义，因此识别篇章单元对的篇章关系效果差。

（3）我们运用严为绒等人[16]的方法，计算待测篇章单元对中的框架对的互信息，选取互信息排序前 4 的框架对，将每一对框架对都和 FRmap 进行对照，得到框架对对应的篇章关系，将在这 4 个篇章关系中出现次数最多的关系判断为待测篇章单元对的篇章关系。在本文语料库上进行测试，所得结果对比如表 9 所示。

表 9 比较结果

识别方法	Ours(核心框架)	Ours(简单配对)	互信息
Acc/%	43.55	32.25	34.65

对比结果显示，运用核心框架进行识别的性能最好。造成这一结果最主要的原因便是本文的语料规模较小，而互信息对语料的依赖性较大。目前，有关中文篇章关系的语料库规模都较小，因此本文的算法对中文篇章关系分析有更大的适用性。

5 总结与展望

本文基于汉语框架语义网识别篇章关系，研究了如何在框架语义层面进行篇章关系的识别。基于汉语框架语义所构建的理论体系中篇章是由与该篇章内容相关的框架集组合而成，因此本文用核心框架代表篇章单元。在识别核心框架过程中，本文用的是最大熵分类模型。在该实验中由于所用语料有限，因此最大的问题便是数据稀疏问题，导致框架的配对中无法包含所有的框架对，在未来的工作中可以对这一方面进行优化，同时有效使用汉语框架语义网的相关资源，如框架的语义角色、框架关系等。

参考文献

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

[2] L Carlson，D Marcu，M E Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory[C]//Proceedings of 2nd SIGdial Workshop on Discourse and Dialogue，2001:1-10.

[3] D Marcu and A Echihabi. An unsupervised approach to recognizing discourse relations[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics(ACL)，2002:368-375.

[4] Prasad R，Dinesh N，Lee A，et al. The Penn Discourse Treebank 2.0[C]//Proceeding of the 6th International Conference on Language Resources and Evalution(LREC)，Marrakech，Morocco，2008:2961-2968.

[5] Emily Piter and Ani Nenkova. Using Syntax to Disambiguate Explicit Discourse Connectives in Text[C]//Proceedings of the ACL-IJCNLP 2009 Conference Short Papers，2009: 13-16.

[6] Lan M，Xu Y，Niu Z Y. Leveraging Synthetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics，2013:476-485.

[7] 孙静，李艳翠，周国栋，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.

[8] 张牧宇，宋原，秦兵，等. 中文篇章级句间语义关系识别[J]. 中文信息学报，2013，27(6):51-57.

[9] 张牧宇，秦兵，刘挺. 中文篇章级句间语义关系体系及标注[J]. 中文信息学报，2014，28(2):28-36.

[10] 姬建辉，张牧宇，秦兵，等. 中文篇章级句间关系自动分析[J]. 江西师范大学学报（自然科学版），2015，39(2):124-131.

[11] 张牧宇，秦兵，刘挺. 中文篇章关系任务分析及语料标注[J]. 智能计算机与应用，2016，6(5):1-4.

[12] 苏娜. 基于框架语义的汉语篇章连贯性研究[D]. 山西大学，2016.

[13] 李茹. 汉语句子框架语义结构分析技术研究[D]. 山西大学，2012.
[14] 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.
[15] 刘挺，车万翔，李正华. 语言技术平台[J]. 中文信息学报，2012，25(6):53-62.
[16] 严为绒，朱珊珊，洪宇，等. 基于框架语义的隐式篇章关系推理[J]. 中文信息学报，2015，29(3):88-99.

作者联系方式：

李国臣，山西省太原市坞城路92号山西大学计算机与信息技术学院，邮编：030006，电话：13903408101，E-mail：lgc1017@163.com；
张雅星，山西省太原市坞城路92号山西大学计算机与信息技术学院，邮编：030006，电话：18234034867，E-mail：1161748628@qq.com，通讯作者；
李茹，山西省太原市坞城路92号山西大学计算机与信息技术学院，邮编：030006，电话：13073538012，E-mail：liru@sxu.edu.cn。