文章编号：1003-0077（2011）00-0000-00

基于功能连接词的隐式篇章关系推理[1]

车婷婷，洪宇，周小佩，严为绒，姚建民，朱巧明

（苏州大学计算机科学与技术学院 自然语言处理重点实验室，江苏省 苏州 215006）

摘要：功能连接词是一种直接表述篇章单元内部语义关系、结构特性和语境发展趋势的词特征。借助功能连接词的这一优势，本文提出一种基于功能连接词的隐式篇章关系推理方法。该方法首先挖掘词级与短语级的功能连接词，划分功能连接词的篇章关系类别；其次，为每个功能连接词构建概念模型，借以描述由功能连接词连接的论元属性，并建立论元概念与篇章关系的映射体系；最后，利用统计策略识别待测论元的概念模型，并借助“概念-关系”映射体系，实现隐式篇章语义关系推理。实验结果显示，本文基于功能连接词构建概念模型的推理方法，相较于现有的基于监督学习的分类方法，系统性能获得显著提升。

关键词：隐式篇章关系；功能连接词；论元概念模型

中图分类号：TP391　　**文献标识码：**A

Predicting Implicit Discourse Relation Based on Functional Connective

Tingting Che, Yu Hong, Xiaopei Zhou, Weirong Yan, Jianmin Yao, Qiaoming Zhu

Key Laboratory of Natural Language Processing of Jiangsu Province

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

Abstract: The functional connective is a word feature that directly expresses interior semantic relations, structure characteristics and the development trend of context of discourse units. With functional connective's advantages, this paper puts forward a kind of methods for predicting relations of implicit discourse based on it. First, this method mines functional connectives of words and phrases grade level, and divides category of discourse relations of functional connectives. Secondly, it can build a concept model for each type of functional connectives to describe argument attributes connected by functional connectives, and establish the mapping system between argument concepts and discourse relations; Finally, using statistical strategy to recognize conceptual model of argument and with "concept-relations" mapping system, it realizes the predictions of the implicit discourse semantic relation. The experimental results show that, the predicting method by constructing concept model based on functional connectives, compared with the existing classification method based on supervised learning, got the significant performance improvements in this paper.

Key words: implicit discourse relation；functional connective word；argument concept model

1 引言

目前，语义分析已从传统的词义、句法研究及句子内的语义角色标注，逐渐深入到语义上连贯且结构上衔接的文本片段（包括子句、句子、段落和篇章）的语义关系研究。篇章语义关系（Discourse Relation）研究旨在推断篇章内部相邻文本片段，或跨度在一定范围内的多个片段间的逻辑关系。篇章语义关系研究不仅能够有效辅助篇章语义的机器学习和篇章组织结构的自动划分，而且在自然语言处理领域有广泛的应用价值：如篇章因果关系可应用于自动问答系统、事件关系抽取和检测（Riaz和Girju[1]，Do和Chen[2]）等；扩展关系可应用于自动文摘生成、篇章关键词识别（王和武等[3]）等；对比关系可应用于情感分析研究，辅助

[1]**收稿日期：**2013-3-20　　**定稿日期：**2013-3-22

基金项目：国家自然科学基金（No. 61003152,61272259,60970056,90920004），教育部博士学科点专项基金（No. 2009321110006, 20103201110021），江苏省自然科学基金（No. BK2011282），江苏省高校自然科学基金重大项目（No. 11KIJ520003）以及苏州市自然科学基金（No. SYG201030, SH201212）

作者简介：车婷婷（1988—），女，硕士生，主要研究方向为自然语言理解，信息检索；洪宇（1978—），通讯作者，tianxianer@gmail.com，男，博士，讲师，主要研究方向为话题检测与跟踪、信息检索等；周小佩（1988—），女，硕士生，主要研究方向为自然语言处理，信息检索。

实现句内的情感极性判断（Zhou和Li等[4]）等。

根据论元（即论述特定语义的文字片段）间是否存在显式连接词，篇章关系分为显式和隐式篇章关系（Explicit & Implicit Discourse Relation）。前者可借助显式连接词（如“因为”）及其关系映射（如“因为”映射为“因果关系”）进行直接的关系检测；后者需根据上下文内容或语义特征，进行间接的关系推理。如（1）中的显式关系可借助连接词“*but*”直接判定为“对比”关系；而（2）本身不具备连接词“*because*”，仅能通过上下文推测为“因果关系”。（注：两例分别抽选自宾州篇章树库PDTB的显式和隐式关系样本集）

（1）**Arg1**[1]**:** She can stay there with no heat
译文：她能够待在不热的地方
Arg2: ***but*** for a parakeet that can be deadly.
*译文：**但**对长尾鹦鹉来说，这可能是致命的*
——Explicit Discourse Relation：Comparison*（显式篇章关系：比较关系）*

（2）**Agr1:** The administration's concerns are understandable
译文： 政府的担心是可以理解的
Arg2: [Implicit="*because*"] the economy is showing signs of weakness.
*译文：**[隐式=“因为”]**经济正呈现疲软现象*
——Implicit Discourse Relation：Contingency*（隐式篇章关系：因果关系）*

PDTB（Penn Discourse TreeBank）语料已针对大量“论元对”（Arg1-Arg2），标注了由显式连接词表征的跨论元显式篇章关系（如（1）），以及包含潜在连接词的跨论元隐式篇章关系（如（2），其中的 “*because*”是由标注者结合具体语义适当添加的）。目前，对于显式篇章关系的研究，精确率已达 93.09%（Pitler 等，2008）[5]。显式关系实例因自身包含连接词，能避免篇章理解歧义；而隐式关系实例欠缺显式连接词等直接线索，须通过上下文、句法、语义信息等自然语言理解的方式进行判断。而上下文信息的不确定性、句子结构的复杂性、语义关系的歧义性以及数据稀疏问题，往往误导关系推理。针对 PDTB 隐式关系检测的最新研究（Wang 和 Su，2010）[6]，第一层关系的分类精度仅能达到 40.0%。

针对隐式篇章关系推理难点，本文经验性发现：篇章中除显式连接词能够直接反映语义关系外，还存在一种潜在反映篇章关系的功能性连接词（Functional Connective，简称 FC）。其与隐式论元对往往构成紧密的语义关系和依存结构，有助于隐式关系推理。如（3）中的功能连接词“*provoke*”非显式连接词，但其潜在地触发了隐式论元对间的因果关系。

（3）**Arg1:** A buildup in inventories can

Arg2: provoke cutbacks in production that can lead to a recession.
*（译文：库存的增加可能**引发**能够导致经济不景气的生产的缩减。）*

（PDTB2.0_Contingency）

根据功能连接词的这一优势，本文提出一种基于功能连接词的隐式篇章关系推理方法。基本思想包括：针对特定篇章关系类别的功能连接词（人工收集183项并划分篇章关系类别），利用大规模语言学资源，挖掘包含这类功能连接词的“论元对”集合，并对这一集合构建概念模型A，形成“概念-关系”映射体系；篇章关系推理过程中，对给定的待测论元对构建概念模型B，并利用统计策略得到与其相似度匹配最高的概念模型A，及其在上述映射体系中对应的篇章关系，实现待测论元对篇章关系的推理。这一过程中，本文利用功能连接词的特性构建概念模型A，用以解决待测论元对概念模型B的稀疏性，完善了推理机制。

本文构建的概念模型，用于描述“同类论元对”或待测论元对的语义特征（注：“同类论元对”即包含一致功能连接词的论元对）。概念模型可细分为实体/行为/状态概念子模型，它们分别为论元对三种属性特征的抽象描述，表征了论元对的语义特征集合及概率分布。

本文组织结构如下：第一章简介隐式篇章关系识别的任务定义；第二章回顾相关工作；第三章给出基于功能连接词推理隐式关系的主体框架；第四章详细介绍功能连接词的挖掘与

[1] 宾州树库（PTB）是对 WSJ 语料进行句法结构标注的公认语料资源 http://www.cis.upenn.edu/~treebank/

分类、概念模型的构建方法、面向“概念-关系”映射的模型匹配方法；第五章给出实验结果并进行分析；第六章进行总结与展望。

2 任务定义

Wang 和 Su 等[6]（2010）定义了篇章关系识别的核心任务，即自动检测同一篇章内，相邻片段（也称论元对）之间的语义关系。隐式篇章关系检测是在没有显式连接词作为推理线索的情况下，对篇章关系予以判定。

PDTB 建立了篇章语义关系体系（Prasad 和 Dinesh 等, 2008）[7]，通用于显式和隐式篇章关系检测，该体系分为三个层级：第一层包含四种主要的关系类别，即比较关系（Comparison）、偶然性关系（Contingency）、扩展关系（Expansion）以及时序关系（Temporal）；第二、三层分别在上一层关系的基础上进一步细分。由此，篇章关系（包括显式和隐式）检测系统的标准输出，即为反映特定论元对篇章语义关系类别的标签（如因果关系）。本文主要针对 PDTB v2 关系体系中第一层的四种隐式篇章关系进行推理分类。

3 相关工作

自 PDTB 和 RSTDT 语料（Carlson 和 Marcu，2001）[8]发布以来，篇章语义分析和篇章结构分析的研究获得了更深层次的发展。目前篇章语义关系识别的研究侧重采用全监督或半监督学习的方法，研究重点在于使用各种语言学特征，实现篇章关系判定和分类。

Marcu 和 Echihabi（2002）[9]使用词对共现特征检测文本片段间隐式篇章关系的存在。Saito 和 Yamamoto 等[10]（2006）在此基础上联合使用短语特征，提升了日文隐式关系检测的性能。Wellner 等（2006）在 GraphBank（Wolf 等，2005）[11]上通过实验证明，显式连接词与论元间的距离特征，对显式关系的整体分析有重要作用，然而在隐式关系检测中无法获得较优性能，主要原因是显隐式关系本身的差异性（隐式论元间不包含显式连接词等）。Pitler 和 Louis 等[12]（2009）首次单独针对 PDTB 中隐式关系进行分类，使用情感词极性、动词短语长度、句子首尾单词对以及上下文等语言特征，最终分类结果优于随机分类的性能。

Soricut 和 Radu[13]（2003）基于 RSTDT 语料，鉴别了不同特征对篇章关系识别的作用，主要验证了单纯的句法特征并不适用于句间的隐式关系识别。Wang 和 Su（2010）[6]基于卷积树核函数提取论元的句法结构特征，第一层隐式关系分类精确率只达到 40.0%。Lin 和 Ng 等[14]（2009）基于全监督学习的分类框架，使用句法结构特征、论元的嵌套关系及成分依存特征（从论元对依存树中抽取常用词汇）等，第二层隐式关系分类精确率达到 40.2%。

Zhou 等[15]（2010）借助预测显式连接词来判断隐式篇章关系，主要通过统计语言模型推测适用于当前隐式论元间的连接词，再将预测的连接词作为附加特征用于分类，篇章关系的四元分类精确率达到 41.35%，而关系的二元分类（即针对四种篇章关系中的某一种，判断待测论元对是否属于这种关系）精确率仅在偶然性和时序关系上有所提升（分别为 70.79% 和 70.51%），但对扩展和比较关系的分类性能仍然偏低，说明通过预测显示连接词推理隐式关系的缺陷。这也是本文选择使用功能连接词，而非显式连接词的原因之一。

4 隐式篇章关系推理框架

本文探究基于功能连接词，构建论元对概念模型，实现隐式篇章关系推理。推理的主体架构主要包括三个方面：基于功能连接词的论元对归类、概念模型的构建与内部聚类和基于“概念-关系”映射体系的隐式篇章关系推理。下面分别予以概述。

4.1 基于功能连接词的论元对归类

本文中对隐式论元对的归类，以及后续“概念-关系”映射体系的构建都需要借助功能连接词。较以往使用 PDTB 显式连接词的研究不同，本文选择功能连接词源于如下因素：

- 相较于功能连接词，显式连接词多为语义不明确的虚词（歧义大）且分布极不均衡，对

论元归类和映射体系的构建往往产生误导。如显示连接词“*since*”同时具有“自从”和“因为”的含义，篇章关系分类需针对性消歧；而“*and*”在论元间的分布概率达0.57（统计自PDTB v2），且很多并不映射为扩展关系，仅表征语气停顿或一致性等。如（4）的篇章关系非“*and*”表征的扩展关系，而是功能连接词“*unlike*”表征的比较关系。

（4）The Cool Athlon is fully supported by AMD, **and** **unlike** an ordinary PC.

*（译文：Cool Athlon电脑全部使用AMD的处理器，这**与**一般的家用电脑**不同**）*

（显式连接词：and - 扩展关系；功能连接词：unlike - 比较关系）

- 多为虚词的显式连接词全局分布极为广泛，使得借助它的论元对归类被极大泛化，无法构建区分不同篇章关系的论元对概念模型。如广泛分布的 “*and*”在构建其关联的论元对概念模型时，将引入大量不同类别的论元对，形成的概念描述不具有显著的语义针对性，其“概念-关系”映射将导致推理过程的盲目性。

因此，本文借助功能连接词和其表征的篇章关系（如4.1节），从TDT4[1]中挖掘包含它们的论元对（通过句法依存弧识别Arg1和Arg2）并归为同类论元对（归类原因为关联相同功能连接词的论元对，内部牵涉到相似的组件知识），并形成论元对与篇章关系的一一映射。其中每个功能连接词对应一类论元对，不按四类篇章关系进行合并。原因是尽管篇章关系类别相同，但不同的功能连接词在连接论元时，往往并不具有绝对一致的适用性。

4.2 概念模型定义

通过抽取功能连接词论元对中与功能连接词有直接依存关系，或待测论元对中依存关系指向较多的三类词集合（依据词性划分实体、行为和状态词），按论元的主被动关系划分为施事词集和受事词集，形成施/受事实体/行为/状态集，以此为基础分别构建施/受事概念子模型，联合形成概念模型。在此过程中，挖掘词集中词特征的相关知识并构造特征向量，并按词集分别聚类（使用ApCluster[16]），每个类簇构成一种概念，且根据聚类来源可标注概念的“归属”（如，施事实体集的类簇归属于施事实体子概念）。

其中，概念模型（包括 A 和 B 两种）可理解为：由施/受事概念子模型构成的，具有不同“归属”标签的概念的集合。如，由“猴”与“猩猩”等词特征形成的类簇，表征了一种“灵长类动物”的概念，归属标签为施/受事实体子概念；由“殴打”和“射击”等词特征形成的类簇，表征了一种“袭击类事件”的概念，归属标签为施/受事行为子概念。

4.3 隐式篇章关系推理

通过获得与待测论元对概念模型 B，映射的功能连接词论元对概念模型 A，及模型 A 关联的功能连接词所对应的篇章关系，以功能连接词为媒介，可形成“概念-关系”映射体系，通过统计经该映射体系输出的最大可能篇章关系，达到推理目的。

5 推理方法详述

本章针对基于功能连接词进行隐式篇章关系推理的方法，分别介绍功能连接词的挖掘与归类、面向论元对概念模型的构建方法和面向“概念-关系”映射的模型匹配方法。

5.1 功能连接词挖掘与分类

对于论元间不包含显式连接词的隐式篇章关系，可通过具有篇章语义连接功能的其它特定词语表现，本文称这类词语为功能连接词。功能连接词是使论元形成特殊语义关系的重要连接机制，对隐式关系的判定、语义分析与推理具有重要作用。可借助功能连接词的语法、语义及依存连接特性，充分挖掘论元间潜在的逻辑关系特征。本文针对 PDTB 第一层四类篇章关系，分别获取了相应的功能连接词（主要为词级与短语级），其对四类隐式篇章关系的表征效果明显。

[1] http://projects.ldc.upenn.edu/TDT4/

表 1 四种隐式篇章关系的功能连接词举例

Tab. 1 Examples of functional connectives for four implicit discourse relations

偶然性关系 （Contingency）	Arg1：That Sea Containers' plan would， Arg2：**result in** shareholders receiving only $36 to $45 a share in cash. *（译文：海上集装箱的计划将会**导致**股东只收到每股 36 至 45 美元的现金分红。）*
扩展关系 （Expansion）	Arg1：Some projections show Mexico importing crude by the end of the century Arg2：**barring** an overhaul of operations. *（译文：一些预测显示墨西哥直到本世界末仍将进口石油，**除非**出现彻底的改革措施。）*
比较关系 （Comparison）	Arg1：It calculates how often the words appear in the story Arg2：**compared with** how often they appear in the entire data base. *（译文：计算此篇故事中词语出现的频率**与**他们出现在整个数据库中的频率**相比较**。）*
时序关系 （Temporal）	Arg1：Federal officials seized the association in April, Arg2：**a day after** the parent corporation entered bankruptcy-law proceedings. *（译文：联邦官员抓住了四月协会的契机，**一天后**控股母公司进入破产法律诉讼议程。）*

表 1 列举了四类篇章关系的功能连接词实例，可以发现对于不包含显式连接词的论元，由于其间功能连接词（如“*result in*”）的存在，可以辅助推理隐式篇章关系。本文通过获取与表 1 中“*result in*”、“*barring*”、“*compared with*”及“*a day after*”类似的功能连接词，从隐式论元本身出发，根据隐式论元的内部联系属性，构建论元概念模型。

表 2 功能连接词举例（未全部列举）

Tab. 2 Examples of functional connectives (not all list)

偶然性关系 （Contingency）	主动词 连词;副/介词	arouse, evoke, provoke；bring about, result in in that case, on the ground(s) that, by reason that; plainly; due to, owing to
扩展关系 （Expansion）	举例；增补 结论；强调	a case in point, as an illustration, such as；what’s more, excluding in a word, to sum up, as far as I know, all in all；emphasis, obviously
比较关系 （Comparison）	对比 让步	in comparison, compared with, best of all, comparatively, alike, except for after all, in spite of, provided /providing (that) , despite
时序关系 （Temporal）	时序	first of all, the next step, after that, at that moment, from now on, in time, all of a sudden, for the time being

本文共收集功能连接词 183 项（如表 2）。其中，表征偶然性关系的 49 项，扩展关系的 84 项，比较关系的 23 项，时序关系的 27 项。每类功能连接词按词性和作用的不同又可细分成小类。四大类功能连接词的收集存在不平衡性，符合自然语言资源中篇章关系分布本身的不平衡性（表 3 列举了 PDTB 中篇章关系的分布情况）。

表 3 PDTB 语料中显式/隐式篇章关系类别分布

Tab. 3 Discourse relation distribution in explicit/implicit classes in the PDTB

篇章关系类别	比较关系 Comparison	偶然性关系 Contingency	时序关系 Temporal	扩展关系 Expansion	关系总数 Total
显式 Explicit (%)	5590 (28.73%)	3741 (19.23%)	3696 (18.99%)	**6431 (33.05%)**	19458
隐式 Implicit (%)	2505 (15.10%)	4261 (25.69%)	950 (5.73%)	**8868 (53.47%)**	16584

5.2 概念模型的构建

本文通过对具有不同“归属”标签的概念的处理，构建概念模型。模型中的每种概念都是其对应特征向量集的聚类类簇，每种概念的形成过程及后续的概念匹配过程，皆需构建特征向量集。即针对表征某类概念的论元进行特征抽取和属性描述。下面分别予以介绍。

- **特征抽取**

特征抽取是结合语言学信息获得论元的关键词及其属性。本文构建概念模型需针对 论元对（功能连接词论元对和待测论元对）进行特征抽取，步骤如表 4 所示。

表 4 特征抽取基本步骤

Tab.4 Feature extraction steps

目标：功能连接词论元对 T1：Arg1 [Functional Connective] Arg2；待测论元对 T2：Arg1 [implicit=?] Arg2 任务：对论元对进行特征抽取
步骤 1：基于所有功能连接词，使用 Lucene 技术从 TDT 大规模新闻语料挖掘关联各个功能连接词的“同类论

元对”；从 PDTB 测试集中提取待测隐式论元对；

步骤 2：利用斯坦福依存分析工具[17]对论元对进行依存分析；

步骤 3：对 T1，选取与功能连接词具有直接依存弧的非停用词（分为施事词和受事词）作为特征词（比如，图 1 中的“*buildup*”，“*cutbacks*”和“*production*”都与功能连接词“*provoke*”存在直接依存弧，它们可可作为 T1 的特征词）；对 T2，选取依存弧指向较多和依存紧密的非停用词作为特征词。

其中，对功能连接词论元对特征抽取使用约束条件的理由为：这类词特征往往与功能连接词存在直接的语义依存，且作为句法主干元素，能够刻画论元的核心含义，有效反映论元间的语义关系。按照语义角色，这类词特征具备“施事”和“受事”以及依据词性划分的“实体”、“行为”和“状态”标签，有助于分类表述论元概念（辅助细粒度的子概念划分），提升概念模型的匹配准确率和基于概念实现关系推理的精度。

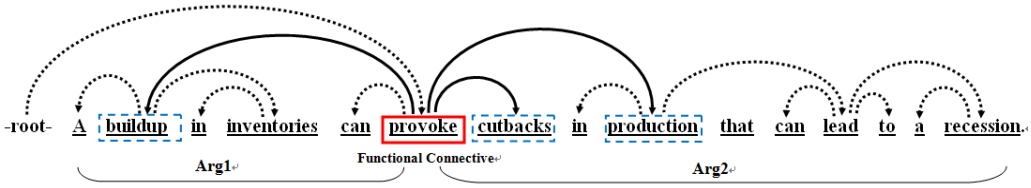

图 1 与功能连接词“*provoke*”关联的论元对的特征抽取

Fig.1 Feature extraction of arguments associated by FC-provoke

- **属性描述**

由于本文论元多为句子级别，篇章长度短，包含的语言学信息不够充分，对经上述步骤抽取的特征词，需要构建其属性向量予以描述，借以扩充特征词的相关属性知识。本文共选取 9 种属性为每个特征词构建属性向量，分别为特征词本身（*Oriword*）、词性（*POS*）、位置（*PL*）、DF 值（*DF*）、与特征词具有单论元（*SingleDep*）和跨论元（*CrossDep*）正/反向依存的非停用词集以及特征词在 WordNet[18]中的同义词（*Syn*）、上位词（*Hype*）和下位词（*Hypo*）。对于第 i 个功能连接词关联的论元对（或待测隐式论元对），其特征词 *Oriword* 的属性向量 $KeyWord_i$ 表述如下：

$$KeyWord_i = \{Oriword，POS，PL，DF，SingleDep，CrossDep，Syn，Hype，Hypo\}$$

POS 特征用来划分该属性向量属于实体、行为还是状态类；*PL* 表示特征词在论元中的绝对位置，Pitler [12]研究证明特殊位置的单词（一般为首尾三个词）具有较强的语义连接功能；*DF* 表示特征词出现在不同类论元对（关联的功能连接词不同）中的频率，DF 小的特征词具有更好的论元对类别区分能力；*SingleDep* 表示与特征词在同一论元中且有依存关系的非停用词集（如图 1 中，特征词“*buildup*”的 *SingleDep* = {*inventories*}）；*CrossDep* 表示与特征词在不同论元中且有依存关系的非停用词集。

- **模型构建**

本文构建的概念模型分为：功能连接词论元对概念模型 A 和待测隐式论元对概念模型 B，两种概念模型通过相似度匹配形成映射关系（如图 2 所示）。

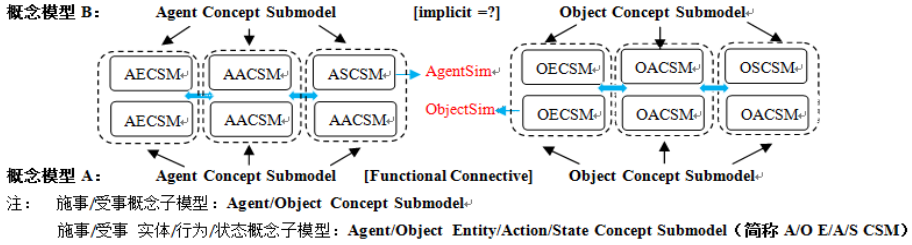

图 2 概念模型的构建

Fig.2 The construction of conceptual model

每种概念模型都由施事和受事概念子模型构成，每种概念子模型中的词特征都按上述“特征抽取”和“属性描述”方法，构造特征向量，按词特征“归属”的词集类别分别聚类

形成概念。施/受事概念子模型正是以此形成的概念的集合。以这种层层细化的方式构造的概念模型，能较系统而全面的描述论元对的语义特征。

5.3 概念模型相似度匹配方法

本文方法框架中重要的一环是将概念模型 B 映射至概念模型 A 的映射体系构建。两种概念模型的映射涉及到两者的相似度匹配问题，其整体相似度度量方法如公式（1）所示：

$$Similarity = \frac{AgentSim + ObjectSim}{2 \cdot e^{|AgentSim - ObjectSim|}} \quad (1)$$

其中，*AgentSim*和*ObjectSim*的含义如图2所示，分别表示施事概念子模型间的相似度和受事概念子模型间的相似度，它们分别是三对概念子模型（施/受事实体概念子模型、施/受事行为概念子模型和施/受事状态概念子模型）的相似度之和。公式中分母起归一化作用。

- **特征向量相似度**

针对*AgentSim*和*ObjectSim*计算过程中提到的三对概念子模型的相似度计算，现以施事实体概念子模型为例，每个概念子模型都由几个类簇构成，每个类簇中的元素都是能表征这一类簇属性的向量（如$KeyWord_i$）。因此一对施事实体概念子模型的相似度，是两组类簇的相似度，即最终细化为类簇中元素的相似度，度量方法如公式（2）所示。

$$Sim(X,Y) = \sum_{i=1}^{N_1} OP_{XY}(i) + \sum_{j=1}^{N_2} Val_{XY}(j) + \sum_{k=1}^{N_3} Set_{XY}(k) \quad (2)$$

其中，*X* 和 *Y* 分别表示需进行相似度计算的两组类簇中的元素（如 $KeyWord_i$ 和 $KeyWord_j$），它们的相似度为 9 维特征的相似度权重之和。因每维特征既有数值形式也有词集合形式，不能直接使用空间向量模型 VSM 计算。公式（2）的第一项为词特征本身（*Oriword*）和词性特征（*POS*）的相似度权重之和（N_1=2）；当 *X* 和 *Y* 的词本身（或词性）特征相同时，$OP_{XY}(i)$ 取 1，否则为 0。公式（2）的第二项为位置（*PL*）和 DF 值（*DF*）特征的相似度权重之和（N_2=2）；计算方法如公式（3）和（4）所示：

$$Val_{XY}(j) = 1 - \frac{|F_j(X) - F_j(Y)|}{Max(S_j(X), S_j(Y))} \quad (3)$$

$$DF = \log\left(\frac{n}{N}\right) \quad (4)$$

公式（3）中，当 j=1 时，$F_j(X)$和 $F_j(Y)$为 *X* 和 *Y* 中的位置特征值；$S_j(X)$和 $S_j(Y)$为构造 *X* 和 *Y* 的论元长度，经归一化后得到 *X* 和 *Y* 的位置特征的相似度权重。当 j=2 时，$F_j(X)$和 $F_j(Y)$为 *X* 和 *Y* 中的 DF 值（计算如公式（4），*n* 为包含特征词的论元类别数，*N* 为论元的类别总数）；$S_j(X)$和 $S_j(Y)$为各自的 *n* 值，经归一化后得到 *X* 和 *Y* 的 DF 值的相似度权重。

公式（2）的第三项为单/跨句依存（*SingleDep*/*CrossDep*）、同义词（*Syn*）和上/下位词（*Hype*/*Hypo*）特征的相似度权重之和（N_3=5），其能有效衡量特征向量间的依存相似度（依存词集交叉词）和背景词汇相似度（同义/上/下位词集交叉词）。计算方法如公式（5）：

$$Set_{XY}(k) = \frac{G(S_k(X), S_k(Y))}{Max(N_k(X), N_k(Y))} \quad (5)$$

公式（5）中 $S_k(X)$和 $S_k(Y)$表示 *X* 和 *Y* 中各自特征词的单句依存词集、跨句依存词集、同义词集、上位词集和下位词集（根据 k 值），$G(S_k(X),S_k(Y))$表示 *X* 和 *Y* 对应的特征词集的词共现数（词集交叉词的个数）；$Max(N_k(X),N_k(Y))$表示 *X* 和 *Y* 各自特征词集的最大长度。

- **概念子模型相似度**

同样以两种概念模型中的施事实体概念子模型（由多个类簇构成）间的相似度计算为例，以特征向量的相似度计算为基础，子模型间的相似度计算即两组类簇间的相似度计算，本文采用三种相似度匹配方法 *CentSim*，*AvgSim* 和 *TopNSim*（如表 5 所示）进行对比实验。

表 5 三种相似度匹配方法

Tab.5 Three kinds of methods about similarity matching

目的：计算两个概念子模型的相似度 *Class1* - 待测论元对的施事实体概念子模型；*Class2* - 功能连接词论元对的施事实体概念子模型	
CentSim：	*Class1* 的中心向量与 *Class2* 的中心向量进行相似度计算
AvgSim：	*Class1* 的中心向量与 *Class2* 的所有向量进行相似度匹配，计算相似度平均值；
TopNSim：	C*lass1* 的中心向量与 *Class2* 中相似度匹配最高的 N 个特征向量相似度平均值。（其中，当 N 为 *Class2* 的向量总数时，即为 AvgSim 判定标准）。

其中，相似度匹配需按照概念的归属进行分类匹配。如，两种论元对对应的施事实体子概念进行匹配，而不能与另一论元对的受事实体子概念或施事行为子概念等匹配。通过统计最优匹配的概念模型 A 所映射的篇章关系（“概念-关系”），推理待测论元对的篇章关系。

6 实验结果与分析

本章给出基于功能连接词推理隐式篇章关系方法的实验结果和评价标准，并通过对比前人利用树核函数和统计语言模型推理的效果，进一步分析本文方法的特点及优越性。

6.1 实验数据评价标准

本文针对 PDTB 第一层四种隐式篇章关系进行推理识别，采用非监督方法，选择 PDTB 中 21-22 章作为测试集。本文对于包含两种或两种以上篇章关系的测试句对，选择最主要的关系类别作为其正确的篇章关系。表 6 列出了测试集中第一层隐式篇章关系的分布：

表 6 测试集中隐式篇章关系的分布

Tab.6 Implicit discourse relations distribution in testing

隐式篇章关系类别	扩展关系 Expansion	偶然性关系 Contingency	比较关系 Comparison	时序关系 Temporal	实例总数 1042
实例个数（比例）	560（**53.74**%）	268（25.72%）	146（14.01%）	68（6.53%）	

本文重现并测试了 Wang 等（2010）[6]基于树核函数抽取句法结构信息，再利用统计策略推理的方法。本文通过与该方法的对比，验证统计建模的可行性。本文也实现了 Zhou 等（2010）[15]在 PDTB 上使用语言模型，构造一致的论元表达模式来预测显式连接词的推理方法，其能与本文构造的功能连接词概念模型推理方法形成很好的对比。为评估推理系统对四种篇章关系的识别性能，本文使用的度量标准如公式 6 所示, 其中，*PosCorrect* 为被正确分为正例的个数，*NegCorrect* 为被正确分为负例的个数，*Sum* 为测试实例总数（1042）。

$$Accuracy = \frac{PosCorrect + NegCorrect}{Sum} \quad (6)$$

6.2 实验结果与分析

- **可行性验证**

本文首次提出利用功能连接词（FC）构建隐式论元对概念模型，与直接表征论元语义关系的显式连接词不同，FC 主要出现在欠缺显式连接词的隐式论元间，通过其语义连接和依存特征，潜在反映隐式篇章关系，这一特点有利于本文在推理隐式关系时加以利用。

本文分析了较高频功能连接词在隐式和显式篇章关系中的分布情况，以验证使用功能连接词作为线索词，构建隐式论元对概念模型的可行性。如图 3 所示，功能连接词在隐式篇章关系中的出现频率较显式更高，尤其高频功能连接词的这一分布差异更为显著（图 3 的小表列举了四种篇章关系中频率最高的功能连接词在显式和隐式篇章关系中的分布情况）。统计结果说明，功能连接词能更好的表征论元间的隐式篇章关系。

然而，尽管功能连接词更多出现于隐式篇章中，但只有较少的待测隐式论元对包含功能连接词（PDTBv2 中 51%的论元间包含 FC），其中真正起到论元间连接作用的功能连接词，出现频率更低（PBTBv2 的 22-23 章中 39%的论元间出现有连接功能的 FC）。因此不能直接通过功能连接词推理待测论元对的隐式篇章关系。本文有效的解决方法是针对性的构建概念模型，以功能连接词为媒介，通过映射和统计的方式推理隐式篇章关系。

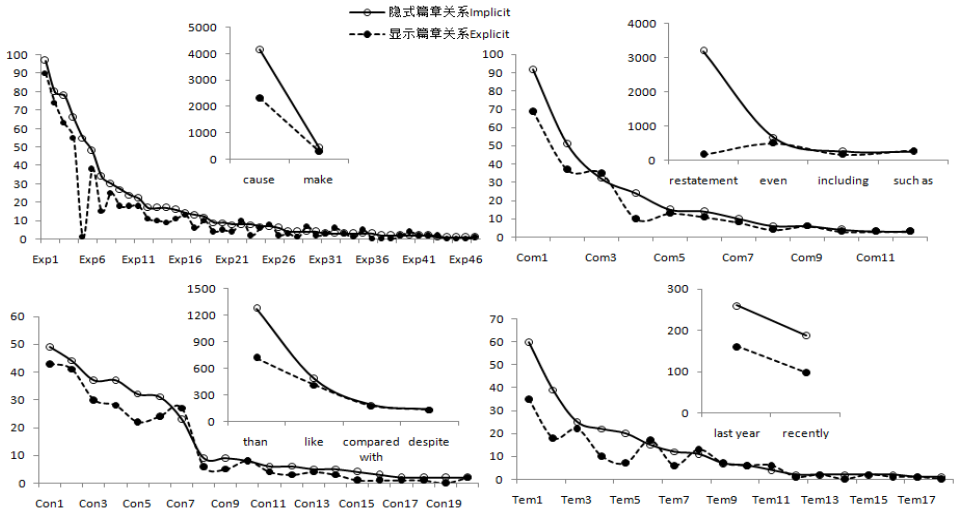

图 3 FC 在显式与隐式篇章关系中的分布情况（纵坐标为 FC 的频率，横坐标为表 2 中 FC 的序列号）

Fig.3 The distributions of FC in implicit and explicit relations（The vertical axis indicates the occurrence frequency of FC, the horizontal axis indicates the sequence number of FC which have been mentioned in Tab.2）

- **相似度匹配方法性能对比**

本文采用三种相似度匹配方法 *CentSim*、*AvgSim* 和 *TopNSim*（如表 5）构建隐式关系推理系统。实验结果对比如表 7 所示，使用平均相似度 *AvgSim* 方法，统计推理隐式关系的精确率最高（53.84%）。而使用聚类中心相似度 *CentSim* 方法，推理系统的精确率最低（50.48%）。

表 7 三种相似度匹配方法的系统性能对比

Tab.7 System performance comparison about three kinds of similarity matching

相似度匹配方法	CentSim	AvgSim	TopNSim
精确率（Accuracy）	50.48%	**53.84%**	53.35%

造成精确率偏差的主要原因是，构建概念模型过程中使用的特征向量，分布较为离散，经过 ApCluster 无指定类别聚类后，类簇的中心向量不能明显表征该类簇中的其它特征向量。而 *CentSim* 方法不能将除中心向量外，有利于篇章推理的其他向量考虑在内。但 *AvgSim* 方法能有效解决中心向量表征类簇效果不好的问题，提高类簇间的相似度匹配性能。

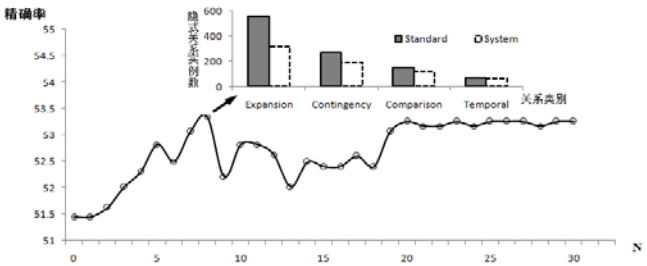

图 4 TopNSim 相似度匹配方法推断隐式篇章关系性能

Fig.4 The predicting implicit discourse relations by TopNSim

另外，实验发现 *TopNSim* 方法中，N 值的变化对系统性能有重要影响。该方法匹配两种概念（类簇）时，将待测类簇的中心向量与候选类簇中相似度最高的 N 个特征向量的相似度权重之和，作为度量两组类簇相似性的标准。如图 4 所示，当 *N*=8 时的系统精确率达到最大值 53.35%。图 4 中的柱形图展示了达到这一精确率时，测试集（*Standard*）和系统判定（*System*）的四种篇章关系实例的分布情况。当 *N* 大于 20 时，精确率近于稳定的原因：一是候选概念中排在较后的特征向量权重较低，累加时对结果影响较小；二是特征向量个数有限，当 N 值增加到足够大时统计结果不再发生变化。

- **推理系统性能比较**

本文将性能最高的系统与 Wang 等（2010）[6]基于树核函数的方法（*Wang_Sys*），以及

Zhou 等（2010）[15]基于非监督语言模型的方法（*Zhou_Sys*）进行对比。此外本文也使用了最大关系类（即所有实例被归类为扩展关系）作为基准（*Baseline*）。表 8 列举了所有系统的精确率，本文方法（*Our_Sys*）在推理隐式篇章关系任务中体现出明显优势，识别精确率较 Wang 和 Zhou 的系统，分别取得 13.84%和 12.49%的性能提升，也高于测试数据中最大类别所占比例。实验结果证实了推理模型构建的正确性以及整体方法的可行性。

表 8 隐式篇章关系推理各方法的性能对比
Tab.8 System performance comparison

系统	识别方法	精确率 Accuracy
Wang_Sys	基于树核函数(Wang)	40.00%
Zhou_Sys	语言模型 (Zhou)	41.35%
Baseline	测试集的最大类别(Expansion)	53.74%
Our_Sys	基于功能连接词的概念模型	**53.84%**

本文系统的性能较 *Wang_Sys* 取得较大提高的原因是，*Wang_Sys* 采用的是基于树核函数抽取句法树中结构化信息，组合句间的时序信息及其它基本特征，进行监督分类的方法。但由于隐式论元对的句法结构较复杂，且仅仅依据篇章中孤立句子的结构信息作为特征来分类显然是不完备的。本文系统性能也优于 *Zhou_Sys* 的原因是，后者通过预测显式连接词，将隐式论元对映射为显式论元对来推理隐式关系，其方法仅基于小规模的显式数据集（PDTB），且仅使用三元语法模型搜索与隐式论元一致的表达模式，严格限制了所构建模式的数量与有效性，使得匹配显式论元对的过程存在缺陷，从而导致预测出的显式连接词不能有效表征隐式论元对的篇章关系。相比之下，本文方法使用丰富的候选资源，从隐式论元本身出发，构建更为完善的概念模型和基于严格相似度度量方法的映射体系，并使用更普遍存在于隐式论元间的功能连接词实现推理。尽管如此，本文工作和 Zhou 的方法在性能上都较优于 Wang 的系统，说明隐式篇章关系识别中模型推理的可行性。且相对简单的映射体系可避免机器学习方法中复杂的语言分析问题，从而减少中间步骤误差引起的错误扩大化现象。

然而，本文的最好性能相较于最大类别的比率仍然较低，Wang 和 Zhou 等的工作甚至远低于最大类别比率，这反映了隐式篇章关系识别难度依然很大，主要是因为隐式关系本身就存在主观性和模糊性，不同的语境下相同的论元对可能形成不同的篇章关系，即使相同的语境下，论元对的语气强度和情感差异也会导致篇章关系的不同。PDTB 语料的 16051 个隐式实例中，有 356 个实例被同时标注多种篇章关系类型；18459 个显式实例中，也存在 532 个同时标注多种篇章关系类型的实例。另外，本文方法中应用的依存分析器的精度，也会影响实验结果。种种现象均表明，隐式篇章关系识别研究将是篇章分析领域的一项既困难同时又富有挑战性的工作。

7 总结与展望

本文首次提出基于功能连接词构建论元概念模型，以无监督的方式实现隐式篇章关系判别。本文利用隐式论元间具有特殊语义连接与依存关系的功能连接词，从隐式关系论元本身出发，提出基于功能连接词构建论元概念模型的篇章关系推理方法。而相关工作中基于复杂语言学特征的监督学习方法，主要是通过利用显式篇章关系特有的属性特征，解决隐式篇章关系的分类问题，忽视了显式与隐式语义关系的本质区别，且复杂的语言学分析会造成中间过程的误差累积，影响最终的分类性能。

另外，本文研究发现目前的隐式篇章关系推理仍存在几大难点问题：1）篇章关系本身存在主观性和模糊性，应充分利用上下文信息辅助隐式篇章关系推理。2）修辞结构在篇章结构中具有重要作用，能有效辅助隐式篇章关系判别，但修辞结构本身就是一项研究难点。

未来工作将借助修辞和情感分析等，扩充现有的功能连接词，进一步挖掘功能连接词的语义特征，并细粒度划分功能连接词的关系类别，完善概念模型的构建方法，进而辅助第一层乃至第二层隐式篇章关系的自动判定。

参 考 文 献

[1] M. Riaz and R. Girju. Another look at causality: Discovering scenario-specific contingency relationships with no supervision[C]//Proceedings of the 4th ICSC, 2010: 361–368

[2] Q. X. Do, Y. S. Chan, and D. Roth. Minimally supervised event causality identification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011: 294–303

[3] 王继成，武港山. 一种篇章结构指导的中文Web文档自动摘要方法[J]. 计算机研究与发展, 2003, 40(3): 398–405

[4] L. Zhou, B. Li, W. Gao, Z. Wei, and K. F. Wong. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011: 162–171

[5] E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. K. Joshi. Easily identifiable discourse relations[C]// Proceedings of the 22nd International Conference on the COLING, 2008: 87-90

[6] W. T. Wang, J. Su, and C. L. Tan. Kernel Based Discourse Relation Recognition with Temporal Ordering Information[C]//Proceedings of the 48th Annual Meeting of the ACL, 2010: 710–719

[7] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. 2008. The Penn Discourse TreeBank 2.0. In Proceedings of the 6th International Conference on LREC 2008, Morocco.

[8] Carlson, L., Marcu, D., and Okurowski, M. E. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Proceedings of the Second SIGDIAL2001, September, Denmark, 1–10.

[9] D.Marcu and A.Echihabi. An Unsupervised Approach to Recognizing Discourse Relations[C]//Proceedings of the 40th Annual Meeting on the ACL, 2002: 368–375

[10] M. Saito, K. Yamamoto, and S. Sekine. Using Phrasal Patterns to Identify Discourse Relations[C]// Proceedings of the Human Language Technology Conference of the NAACL, 2006: 133–136

[11] Wolf, F., and Gibson, E. 2005. Representing discourse coherence: a corpus-based analysis. In Proceedings of the 20th International Conference on the COLING, Morristown, NJ, USA, 134–140.

[12] E. Pitler, A. Louis, and A. Nenkova. Automatic Sense Prediction for Implicit Discourse Relations in Text[C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009,Volume 2: 683–691

[13] R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information[C]// Proceedings of the HLT/NAACL, 2003: 149-156

[14] Z. Lin, H. T. Ng, and M. Y. Kan. Automatically Evaluating Text Coherence Using Discourse Relations[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, Volume 1: 997–1006

[15] Z. M. Zhou, Y. Xu, Z. Y. Niu, M. Lan, J. Su, and C. L. Tan. Predicting Discourse Connectives for Implicit Discourse Relation Recognition[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010: 1507–1514

[16] http://www.bioinf.jku.at/software/apcluster/

[17] http://nlp.stanford.edu/software/lex-parser.shtml

[18] E.Pitler and A.Nenkova. Revisiting readability: A unified framework for predicting text quality[C]//Proceedings of the Conference on the EMNLP, 2008: 186–195

作者联系方式：车婷婷　**地址：**江苏省苏州市苏州大学本部 十梓街 1 号 205 信箱转理工楼 416 室　**邮编：**215006　**电话：**15952418885　**电子邮箱：**chetingting1101@gmail.com