北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.027

汉语隐式篇章关系识别

孙静[1] 李艳翠[1,2] 周国栋[1,†] 冯文贺[3]

1. 苏州大学计算机科学与技术学院，苏州 215006; 2. 河南科技学院信息工程学院，新乡 453003;
3. 河南科技学院人文学院，新乡 453003; † 通信作者, E-mail: gdzhou@suda.edu.cn

摘要 采用一个自建的汉语篇章结构语料库(隐式关系占 80%)进行隐式关系识别。语料中将篇章关系分成三个层次，其中第一层包含因果、并列、转折、解说 4 大类。在此语料上，采用上下文特征、词汇特征、依存树特征，利用最大熵的分类方法对 4 大类关系进行识别。实验结果显示，总正确率为 62.15%，其中并列类识别效果最好, F1 值达到 75.26%。

关键词 篇章结构分析；篇章关系；隐式关系识别；汉语篇章语料库

中图分类号 TP391

Research of Chinese Implicit Discourse Relation Recognition

SUN Jing[1], LI Yancui[1,2], ZHOU Guodong[1,†], FENG Wenhe[3]

1. Department of Computer Science and Technology, Soochow University, Suzhou 215006; 2. School of Information Engineering, Henan Institute of Science and Technology, Xinxiang 453003; 3. School of humanities, Henan Institute of Science and Technology, Xinxiang 453003;
† Corresponding author, E-mail: gdzhou@suda.edu.cn

Abstract The authors use a self-built Chinese Discourse Treebank(80% relations are implicit) to recognize implicit relations. In this corpus, discourse relations are divided into three layers, the first one has four types: causality, coordination, transition and explanation. Based on this corpus, maximum entropy classifier is employed to identify four types relations with context feature, lexical feature and dependency parse feature. Experimental results show that total accuracy rate is 62.15% and the identification effect of coordination is the best, F1 value reaches 75.26%.

Key words discourse parsing; discourse relation; implicit relation recognition; Chinese Discourse Treebank

随着字、词、短语、句子级别研究的逐渐深入和成熟，越来越多的研究者把研究重点转向篇章层级。篇章有时也称语篇或话语，通常指由一系列连续的子句、句子或句群构成的，有意义、传达一个完整信息、前后衔接、语义连贯的语言整体单位。在一个篇章中，各子句之间并不是杂乱无章的堆放在一起，而是具有一定的层次结构和语义关系，只有分析出其中的层次结构及语义关系，才能对篇章进行深入分析和理解。篇章结构分析是自然语言处理的一个核心问题，也是近几年的一个研究热点和难点。篇章结构分析在自动文摘[1−2]、问答系统[3−4]、指代消解[5]和篇章连贯性评价[6]等方面都有所应用。

篇章关系是指同一篇章内部，相邻片段或跨度在一定范围内的两个片段之间的语义连接关系，如条件关系、转折关系、因果关系等[7]。近年来，英语篇章关系识别研究增长迅速，其中一个原因是修辞结构理论篇章树库(Rhetorical Structure Theory-Discourse Treebank, RST-DT)[9]、宾州篇章树库(Penn Discourse TreeBank, PDTB)[10]和图库(GraphBank)[11]等大规模人工标注的篇章语料库的出现，为研究者提供了有价值的资源和平台。

相对于英语篇章关系研究的快速发展，汉语的研究还很少，其中影响其发展的一个重要原因是影

863 计划(2012AA011102)、国家自然科学基金(61273320)和教育部人文社科青年基金(13YJC740022)资助
收稿日期: 2013−06−22; 修回日期: 2013−09−30; 网络出版时间: 2013-11-11 10:25

响力高的大规模语料库的缺乏。本文使用的语料库是自建的汉语语料库——Chinese Discourse Treebank, CNDB。

篇章关系识别作为篇章结构分析的重要内容之一, 可以分为显式关系识别和隐式关系识别。显式关系识别是指在给定连接词(如例 1, 下划线标注连接词“因此”)的情况下, 判断两个论元之间存在何种逻辑关系。因为大多数的情况下, 连接词都是非歧义的, 所以显式关系识别容易实现, 可以仅根据连接词本身就能比较准确地识别论元间存在的特定逻辑关系。英语中通常可以达到 94%左右的 *F*1 值(*F*1-measure)[8], 汉语方面, 在本文的实验中, 可以达到 91.49%的总正确率。隐式关系识别由于连接词缺失(如例 2, 缺失方括号中连接词“因此”), 判断两个论元之间存在何种逻辑关系较困难, 通常只能根据一些语言学特征进行关系的识别, 一般准确率不高。并且根据统计在本文使用的汉语语料库,其隐式关系占有很高的比重, 这说明隐式关系识别的重要性、困难性和挑战性。

例 1 浦东开发开放是一项振兴上海, 建设现代化经济、贸易、金融中心的跨世纪工程, 因此大量出现的是以前不曾遇到过的新情况、新问题。

例 2 他在两起汽车走私案中触犯刑律, [因此]构成走私罪。

本文利用目前已标注的小规模 CNDB 进行汉语隐式篇章关系识别, 应用了上下文特征、词汇特征、依存树特征, 总正确率达到 62.15%。

1 相关工作

根据是否存在连接词, 篇章关系识别可以分为显式篇章关系识别和隐式篇章关系识别两大类。

Marcu 等[12]把显式关系中的连接词去掉, 得到隐式关系语料库, 在这个语料上, 采用词对信息, 用贝叶斯模型进行了关系的识别, 实验结果证明了词对信息的有效性。Pitler 等[13]给出一系列实验自动识别隐式篇章关系, 使用了多个语言学的特征, 包括极性信息、货币/百分比/数字、词的形态、动词短语长度、两个论元的前 3 个单词及相关组合(First-Last-First3)等, 所识别的隐式关系主要是 4 大类, 分析了各种特征对各类关系的作用, 最后得出使用最有效的特征分类。Lin 等[14]应用词对信息、上下文信息、句法信息和依存信息识别 PDTB 中的二级隐式篇章关系, 正确率为 40.2%, 比当时报告的最好结果高 14.1%。Huang 等[15]进行汉语篇章关系识别, 先构建了一个篇章语料库, 从 Sinica Treebank 3.1 中随机抽取 81 篇石油和旅游领域文档进行标注, 完成了 3081 个句对的小规模的中文篇章树库, 在标注过程中简单的把句子作为一个论元, 标注了 4 种关系(temporal, contingency, comparison, expansion), 然后利用词、词性和上位词等特征训练分类器。

考虑到词、词性、词对信息、上下文信息和依存信息在已有工作中的有效性, 本文拟将此应用至汉语隐式篇章关系识别中, 探讨汉语的隐式篇章关系识别问题。

2 CNDB 简介

CNDB 采用树的形式表示汉语的篇章结构, 每一段落构建一棵篇章结构树。例 3 给出一个标注实例, 图 1 是例 3 的一棵篇章树。

例 3 中给出了和关系相关的部分标注信息。“<R”开头的每一行表示一个关系, 其中“ConnectiveType”表示是显式关系还是隐式关系, “Connective”给出关系中的连接词, “RelationType”给出关系类型, “ConnectiveAttribute”表示显式关系中连接词是否可删或者隐式关系中是否可添加连接词, “Sentence”中用“|”分割表示出每个关系的论元, “SentencePosition”给出论元在篇章中的位置, “ChildList”表示包含的孩子关系。如例 3 中的<R ID=“2”表示的隐式关系属于“例证关系”类别, 可以添加连接词“例如”, 包括两个论元: “浦东开发开放……新情况、新问题。对此, 浦东不是简单的采取……纳入法制轨道。”和“去年初……没有发现一例回扣。”, 这两个论元在篇章中的位置分别为: “1…167(表示从第一个字到一百六十七个字)”和“168…230”, 此关系中还包括以<R ID="3"和<R ID="8"开始的两个子关系。语料中还包含关系主次等信息, 这里不再一一列举。

例 3:

<P ID="3">

<R ID="2" …… ConnectiveType="隐式关系"…… Connective="例如" RelationType="例证关系" ……ConnectiveAttribute=" 可 添 加 " ……Sentence="浦东开发开放……新情况、新问题。对此, 浦东不是简单的采取……纳入法制轨道。|去年初……没有发现一例回扣。"

SentencePosition="1…167|168…230"
ChildList="3|8" ……/>
<R ID="3" …… ConnectiveType="显式关系"…… Connective="对此" RelationType="条件关系" ……ConnectiveAttribute="不可删除" …… Sentence="浦东开发开放……新情况、新问题。|对此，浦东不是……纳入法制轨道。" SentencePosition="1…60|61…167"ChildList="4|5" ……/>
……
</P>

例 3 中共有 3 个完整的句子，每个句子内部又有多个基本篇章单位，本文称为子句(Clause)，含传统单句及复句中的分句，文献[16]详细阐述了子句的定义方法并给出简单的子句识别结果。结构上，子句至少包含一个谓语部分，至少表达一个命题；功能上，子句对外不作为其它子句结构的语法成分，子句和子句间发生命题关系；形式上，子句间一定有标点分割，通常是逗号、分号和句号等。

图 1 中的叶子节点的字母标记表示基本篇章单位，中间节点表示篇章关系，各基本篇章单位根据其篇章关系组合后形成高一级篇章单位，进而通过再组合形成更高一级篇章单位，如此层层组合，最后形成一棵篇章结构树。

篇章关系根据是否有连接词可分为显式关系和隐式关系，同时人为将篇章关系分成 3 个意义层次，第一层包括 4 大类：因果类、并列类、转折类和解说类。4 大类下面还包含第 2 层，例如因果类包括：因果关系、推断关系、假设关系、目的关系、条件关系和背景关系。第 2 层下面包含第 3 层，第 3 层为连接词，例如表示因果关系的连接词有：因此、由此、以此、所以、因而和因为等，具体见表 1，其中第 3 层仅显示部分连接词。图 1 中的中间节点就是用 3 层结构表示的篇章关系节点，“转折类：转折关系：不是…而是”表示显式关系，“因果类：目的关系：(以)”表示隐式关系，括号中的词语表示可添加的连接词。例 4 和例 5 给出关系实例(斜体代表 Arg_1，粗体代表 Arg_2，下划线表示连接词，各实例后的括号内容为对应的层级结构)。

例 4 [显式关系]

*对此，浦东不是简单的采取“干一段时间，等积累了经验以后再制定法规条例”的做法，*而是**借鉴发达国家和深圳等特区的经验教训，聘请国内外有关专家学者，积极、及时地制定和推出法规性文件，使这些经济活动一出现就被纳入法制轨道**。(转折类：转折关系：不是…而是)

例 5 [隐式关系]

*而是借鉴发达国家和深圳等特区的经验教训，聘请国内外有关专家学者，积极、及时地制定和推出法规性文件，***使这些经济活动一出现就被纳入法制轨道**。(因果类：目的关系：(以))

目前已经标注了 CTB6.0 中的 158 个文档(chtb001–chtb0130, chtb0211–chtb0240)，CNDB 目前有 739 段，每段平均 3 个句子。总共有 3398 个子句 (叶子节点)，2331 非终节点 (一个非终节点表示一个关系，二元或多元关系)。在 CNDB 中，显式关系有 453 个，隐式关系有 1878 个，达到 80%。而宾州篇章语料库中隐式篇章关系与显式篇章关系大约各占一半[17]，相对于英语来说，汉语中隐式关系的识别更具有挑战性，更加困难，其必要性和重要性也不言而喻。

标注语料由两组标注者共同完成，为便于一致性分析，本语料的 chtb041–chtb0100 这 60 篇文档让两组标注者分别进行标注，根据标注结果计算得到：CNDB 子句分割标注一致性为 94.6%，Kappa 值[18]

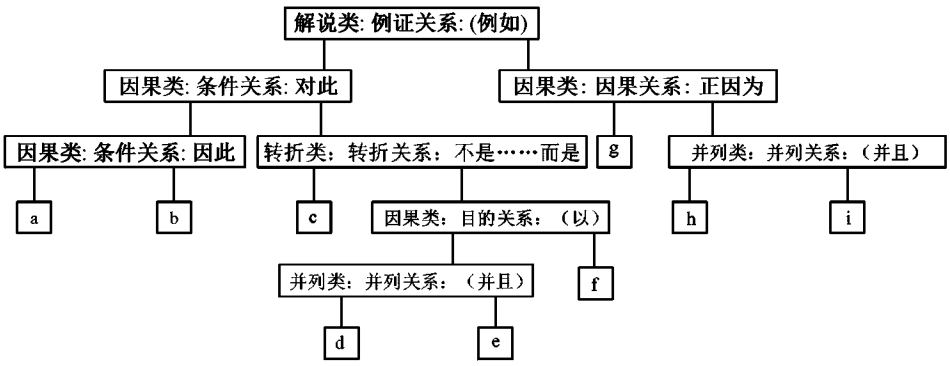


图 1 例 3 的篇章关系树实例

Fig. 1 An instance of example 3's Discourse Relation Tree

为 0.891；显式关系和隐式关系判断一致性为 95.9%, Kappa 值为 0.954。

下面将本语料与两个已有的影响较大的英语篇章结构语料——修辞结构理论篇章树库(RSTDT)和宾州篇章树库(PDTB)做简单对比。

在基本篇章单位的定义上，本文的基本篇章单位(子句)一定有标点作为标志，一般是小于或等于句子的单位。RSTDT 中的基本篇章单位可以小到短语, PDTB 中论元可以大到多个句子，小到从句。在连接词的处理上，RSTDT 没有标注连接词，本文和 PDTB 都标注了连接词。在篇章关系的处理上，RSTDT 和 PDTB 都标注了关系类别，本文将关系和连接词区分开，以使篇章结构分析结果适用于不同任务。在结构树表示上，RSTDT 和本文均可构建完整的篇章结构树，PDTB 则没有着意构建篇章结构树，但可以根据已有关系推导出部分结构树。

3 方法

本文仅针对 CNDB 中第一层的 4 大类隐式篇章语义关系进行识别，并采用一种基于最大熵模型的有监督自动识别方法。最大熵分类器采用张乐编写的 Maxent 机器学习工具包①。我们首先从 CNDB 的两个论元对中抽取出词汇、上下文信息及基于依存树结构的结构化特征，用来训练最大熵分类器，最后预测测试实例的两个论元所具有的篇章关系类别。本文用 chtb001-chtb0130 中的 625 个段落作为训练集, chtb0211-chtb0240 中的 114 个段落作为测试集。表 1 给出训练集和测试集中隐式篇章关系的第一层级的分布情况。从表 1 可知，训练实例 1641 个，测试实例 237 个，其中并列类所占比例最大，转折类只有 17 个训练实例, 7 个测试实例。

3.1 特征提取

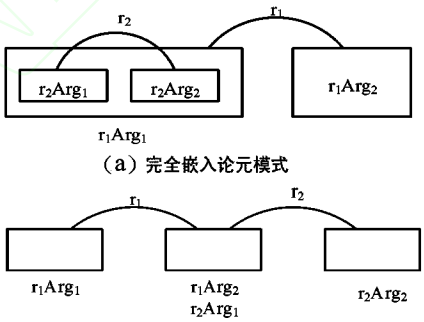
借鉴 Lin 等[14]在 PDTB 上进行隐式关系识别时所用的特征，结合问题本身的特点，本文提取下面 3 组特征。

3.1.1 上下文特征

CNDB 中，每一段话构成一• 篇章结构树，在段落中，每一行都表示一个关系，关系中一般包括两个论元(Arg_1, Arg_2)，但有时也包括多个论元，如例 6 所示(用“|”分割)。通过观察语料库发现，这种情况绝大多数出现在并列关系中，本文认为 a 和 b 是并列关系，b 和 c 同样也是并列关系，因此，把例 6 表示成两个论元不同的并列关系，其一的论元为 a, b，其二的论元为 b, c。

在语料中，一个隐式篇章关系可能包括多于一个的关系类别，如例 7 所示。这种情况下，本文把它表示成两种具有不同类别的关系。

例 6 Sentence="[a]改革开放以来，崇明县的经济建设和对外开放发展迅猛，|[b]外商投资企业不断



（a）完全嵌入论元模式

（b）共享论元模式

图 2 两种篇章依赖结构类型

Fig. 2 Two types of discourse dependency structures

表 1 语料中隐式关系类别分布

Table 1 Distribution of implicit relations of CNDB

第一层	第二层	第三层	训练实例(所占比重/%)	测试实例(所占比重/%)
因果类	因果关系；推断关系；假设关系；目的关系；条件关系；背景关系	因此，所以，因为，由于…所以，如果…那么	297(18.10)	67(28.27)
转折类	转折关系；让步关系	虽然…但，即使…也	17(1.04)	7(2.95)
并列类	并列关系；顺承关系；递进关系；选择关系；对比关系	同时，并，并且，然后	979(59.66)	123(51.90)
解说类	解说关系；总分关系；例证关系；评价关系	总之，例如	348(21.2)	40(16.88)
总和			1,641	237

① http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html#faq

增多, |[c]进出口货物大量增加, "。

例 7 *所有境内机构借用国际商业贷款应当经国家外汇管理局批准*。**未经批准而擅自签订的国际商业贷款协议无效，外汇局将不予办理外债登记，银行不得为其开立外债专用帐户，借款本息不得擅自汇出**(转折类、并列类)。

观察处理后的语料可以发现，篇章关系对之间存在着两种最普遍的模式：完全嵌入论元模式和共享论元模式。如图 2(a)所示的完全嵌入论元模式，篇章关系 r_2 的两个论元 r_2Arg_1 和 r_2Arg_2 完全被关系 r_1 的论元 r_1Arg_1 包含；如图 2(b)所示的共享论元模式，篇章关系 r_1 和 r_2 共享一个论元。

借鉴 Lin 等的上下文特征思想，根据本文中具体的语料现象，给出两种篇章模式(完全嵌入论元模式和共享论元模式)的 6 个特征，如下所示。curr 指当前关系, pre 指上一个关系, next 指下一个关系。其中完全嵌入论元模式与 Lin 等[14]的不同，本文判断当前关系是否完全嵌入上一个关系的 Arg_1 或者 Arg_2 中，下一个关系是否完全嵌入到当前关系的 Arg_1 或者 Arg_2 中。当上一个关系或者下一个关系为显式关系时，用它们的连接词作为上下文特征，记为 FCon。

完全嵌入论元模式：	共享论元模式：
curr embedded in pre.Arg_1	prev.Arg_2=curr.Arg_1
curr embedded in pre.Arg_2	curr.Arg_2=next.Arg_1
next embedded in curr.Arg_1	
next embedded in curr.Arg_2	

3.1.2 词汇特征

词对特征：在英语篇章关系处理中，已有实验证明，词对特征在关系的识别中非常有效[11]。在本文中也把此特征加入实验，(w_1, w_2)，w_1 指 Arg_1 中一个词，w_2 指 Arg_2 中一个词。此特征记为 FWP。

词和词性：我们认为句子中的动词在一定程度上能够反应出句子的意向，因此判断论元是否有以下的词性标注：“VV”，“VC”，“VE”，“VA”，“CS”，“CC”，“AD”，“DEV”，“BA”，“SB”，“LC”，如有，则给出该词性对应的词和词性及其组合。此特征记为 FVwp。

3.1.3 依存树特征

依存树描述各个词语之间的依存关系，即指出词语之间在句法上的搭配关系，这种搭配关系是与语义相关联的。借鉴 Lin 等的特征，首先利用 Stanford 句法分析器对每个句子进行依存句法分析，然后从每个论元对应的依存树中选择所有拥有被支配者的词和依存类型。每个论元根据其跨度的不同可能对应着完整依存树、子树或者多棵树。

图 3 给出例 1 的完整依存树，对于整棵树可以抽取如下的依存规则：是_top_attr_dep、开放_nn_nn、工程_clf_rcmod_amod、是_advmod_top_attr、振兴_dobj_conj_cpm，以此类推，遍历整个依存树。核对每一个规则是否出现在 Arg_1 中、Arg_2 中或同时出现在两者中，最后表示为 3 个二元特征。此特征记为 FDep。

4 实验结果及分析

采用第 3 节提出的全部特征生成隐式篇章关系对应的特征实例集，然后应用张乐的最大熵工具包进行实验，得到标注的总正确率为 62.15%，表 2，3 和 4 给出了不同的实验结果。实验设置所有关系为

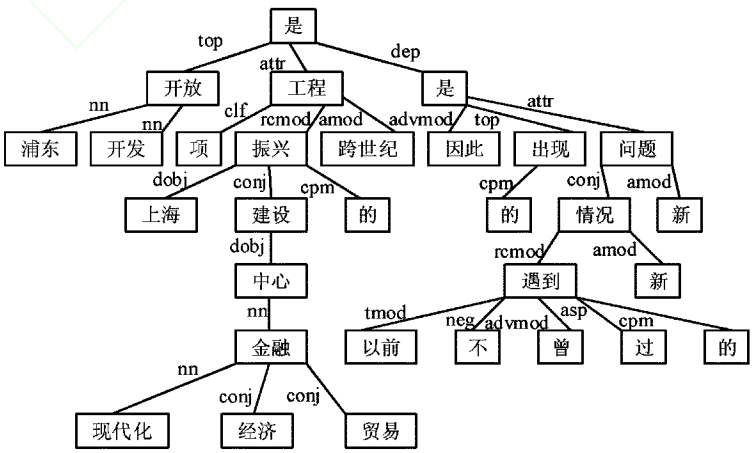


图 3 例 1 的依存分析树

Fig. 3 A dependency tree of example 1

并列类时为基准系统，正确率为 51.9%。

在实验中，词汇特征中的词和词性都采用 CTB6.0 中给出的标准分词和词性标注结果，而依存树是在标准分词的情况下使用 stanford[②]工具自动得到的。

表 2 给出单个特征的总正确率，然后给出所有特征组合(FAll)的总正确率，通过观察表 2 中的实验结果，可以看出本文中选择的几组特征都是有效的，总正确率都超过了基准系统。其中的 FVwp 特征只是简单的判断论元中的动词、副词和连词等，取其词和词性，总正确率达到 54.17%，而其中上下文特征的识别效果最好，达到 60.76%，这与语料的结构有关。

为了检验组合特征对实验结果的影响，逐步累加特征，得到表 3 中的结果。通过表 3 可以看出组合特征的识别结果要优于单个特征的结果。其中，词、词性，依存特征词对特征和上下文特征的组合的效果是最优的，证明本文选择的特征是效的。

表 4 给出在所有特征的组合条件下的每种类别的 *P*, *R* 和 *F* 值，然后给出微平均结果，通过观察表 4 可以看出转折类没有被识别出来，原因是数据的稀疏问题，训练语料中只有 17 个实例，仅占 1.04%。并列类的识别效果最好，从表 1 中关系的分布可以看出并列类占有的比重最高(59.66%)，而且上下文特征中的共享论元模式大多数从并列类中得到，对于并列类识别有针对性，所以并列类的识别最优，在实验中很容易把其他关系误判为并列类，如“但他却无法用“跳”来表达自己的激动之情。|三岁时一场高烧，使他患上了严重的小儿麻痹后遗症，这一年他被福利院收养。”这两句之间本是因果类，但在实验中，却被归为并列类。解说类的识别结果次于并列类，因果类再次之，但是与并列类的识别结果差距很大。下一步需要继续研究，提出针对解说类和因果类行之有效的特征。

表 2　单个特征及所有特征的总正确率

Table 2　Accuracy of individual feature and all fetures

Feature	Word& pos	Dependency rules	Wordpairs	Context	Acc./%
FVwp	+	–	–	–	54.17
FDep	–	+	–	–	54.51
FWP	–	–	+	–	56.25
FCon	–	–	–	+	60.76
FAll	+	+	+	+	62.15

表 3　不同特征组的总正确率

Table 3　Accuracy of different feature groups

Feature	Word&pos	Wordpairs	Dependency rules	Context	Acc.(%)
FVwp	+	–	–	–	54.17
FWP	+	+	–	–	58.68
FDep	+	+	+	–	61.11
FCon	+	+	+	+	62.15

表 4　4 大类别 Precision, Recall 和 F1-measure, "–"代表 0.00

Table 4　Accuracy of different feature groups

类别	*P*%	*R*%	*F*%
因果类	50.0	5.88	10.34
并列类	62.78	95.43	75.26
解说类	47.06	19.51	27.18
转折类	–	–	–
All(微平均)	39.96	30.20	28.20

5　结束语

本文研究了基于汉语篇章语料库(CNDB)中的 4 大类别(因果类、并列类、解说类和转折类)的隐式篇章关系识别问题。相对于英语篇章结构分析的快速发展，汉语方面的进展还比较缓慢，相关工作比较少，一个重要原因是缺少有影响力的大规模的语料库。本文给出一个小规模的自建的汉语篇章语料库(CNDB)，应用词汇特征、上下文特征和依存树特征对隐式关系进行了简单探讨，实验结果表明方法可行。

通过本文使用的语料库可以看出，隐式关系占有很高的比重，又因为隐式关系中缺乏连接词，导致汉语隐式篇章关系识别具有很大的挑战性。本文进行了简单探讨，在以后还有很多工作需要去做，如提出针对汉语特点的特征等。

参考文献

[1] Marcu D. The theory and practice of discourse parsing and summarization. The MIT press, 2000

[2] Louis A, Joshi A, Nenkova A. Discourse indicators for content selection in summarization // Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for

② http://nlp.stanford.edu/software/lex-parser.shtml#Download

Computational Linguistics, 2010: 147–156

[3] Verberne S, Boves L, Oostdijk N, et al. Discourse-based answering of why-questions. Traitement Automatique des Langues, Discourse et document: traitements automatiques, 2007, 47(2): 21–41

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

[5] Webber B, Stone M, Joshi A, et al. Anaphora and discourse structure. Computational Linguistics, 2003, 29(4): 545–587

[6] Lin Z H, Ng H T, Kan M Y. Automatically evaluating text coherence using discourse relations // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT). Portland, OR, 2011: 997–1006

[7] 周小佩, 洪宇, 车婷婷, 等. 一种无指导的隐式篇章关系推理方法研究. 中文信息学报, 2012, 2: 3

[8] Pitler E, Nenkova A. Using syntax to disambiguate explicit discourse connectives in text // Proceedings of the ACL-IJCNLP 2009. Stroudsburg: Association for Computational Linguistics, 2009: 13–16

[9] Carlson L, Marcu D, Okurowski M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory // Proceedings of the SIGDIAL. Stroudsburg: Association for Computational Linguiscs, 2001: 1–10

[10] Prasad R, Miltsakaki E, Dinesh N, et al. The penn discourse treebank 2.0 annotation manual. Technical Report, IRCS-08-01. USA: University of Pennsylvania, 2008

[11] Wolf F, Gibson E. Representing discourse coherence: a corpus-based analysis. Journal of Computational Linguistics, 2005, 31(2): 249–288

[12] Marcu D, Echihabi A. An unsupervised approach to recognizing discourse relations // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Morristown, NJ, 2002: 368–375

[13] Pitler M, Louis A, Nenkova A. Automatic sense prediction for implicit discourse relations in Text // Proceedings of the 47th annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 2009: 683–691

[14] Lin Z H, Kan M Y, Ng H T. Recognizing implicit discourse relations in the penn discourse treebank // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009: 343–351

[15] Huang H H, Chen H H. Chinese discourse relation recognition // Proceedings of the IJCNLP 2011. Chiang Mai: Natural Language Processing of the Asian Federation, 2011: 1442–1446

[16] 李艳翠, 冯文贺, 周国栋, 等. 基于逗号的汉语子句识别研究. 北京大学学报: 自然科学版, 2013, 49(1): 7–14

[17] Pitler E, Raghupathy M, Mehta H, et al. Easily identifiable discourse relations. Technical Reports (CIS), 2008: 85-88

[18] Sim J, Wright C C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy, 2005, 85(3): 257–268