汉语显式篇章关系分析*

丁 彬，孔 芳[1]，李 生，周国栋

（苏州大学 计算机科学与技术学院，江苏省 苏州市 215006）

摘要：篇章关系分为显式和隐式两种。显式关系的显著特征是篇章的基本单元之间存在显式连接词。针对汉语显式篇章关系，构建了包括汉语连接词识别和篇章关系分类的显式篇章关系分析平台。选取汉语宾州树库（Chinese Penn Treebank, CTB）中的 500 篇文本进行了汉语显式篇章关系标注；结合连接词的中心词，采用最大熵分类器构建了汉语连接词识别模块，其性能 F_1 值达到了 66.79%；基于连接词及其词性等上下文特征，构建了篇章关系分类器，其在最顶层四大类语义关系上的分类性能的 F_1 值为 91.92%。

关键词：连接词识别；语义关系分类；最大熵分类器

中图分类号：TP391　　**文献标识码：**A

Explicit Discourse Relation Parsing Of Chinese Text

Bin Ding, Fang Kong, Sheng Li , Guodong Zhou

(School of Computer Sciences and Technology, Soochow University, Suzhou, Jiangsu, 215006, China)

Abstract: Discourse relations can be expressed explicitly or implicitly. This paper focused on explicit discourse relations that are explicitly signaled by discourse connectives. We proposed an explicit discourse relation parsing platform, containing connective identification and sense classification. Using 500 texts from the Chinese Discourse TreeBank corpus (CTB), we annotated an explicit discourse relations corpus. Considering headwords of connectives, we constructed a connective identifier using maximum entropy based on this corpus, which resports F_1 of 66.79%. And a sense classifier based on the context of connective itself is proposed and reports F_1 of 91.92%.

Keywords: connectives identification; sense classification; maximum entropy classifier

1 引言

篇章是指由一系列连续的从句、复句或句群构成，传达一个完整信息、前后衔接、语义连贯的语言单位。篇章分析的主要任务包括研究篇章的内在结构，理解文本单元间承接的语义关系等。篇章分析是自然语言领域至关重要的一部分，对自然语言处理的许多应用，例如问答系统、指代消解和篇章连贯性评价等有着重要的作用。

近年来，随着宾州篇章树库（Penn Discourse TreeBank, PDTB）的发布，英文篇章分析越来越受到关注，许多基于它的研究工作陆续展开。

本文借鉴 PTB 和 RST 英文篇章标注体系，选取汉语树库（Chinese Treebank, CTB）中的 500 篇文本进行了汉语显式篇章关系的标注，并基于这一语料分析了词法和句法特征对汉语显式篇章关系的作用。

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然基金重点项目（61333018）；国家 863 项目（2012AA011102）；国家自然基金项目（61273320）

作者简介：丁彬（1991——），女，硕士研究生，主要研究方向：自然语言处理，篇章处理。Email: 20124227006@suda.edu.cn；孔芳（通讯作者，1977——），女，副教授，主要研究方向：自然语言处理，指代消解，篇章处理。Email: kongfang@suda.edu.cn；李生（1989——），男，硕士研究生，主要研究方向：自然语言处理，篇章处理。Email:qcl6355@gmail.com；周国栋（1967——），男，教授，主要研究方向：自然语言处理。Email: gdzhou@suda.edu.cn。

本文组织如下：第 2 节介绍了显式篇章分析的相关工作；第 3 节介绍了汉语显式篇章关系语料；第 4 节给出了一个基于词法和句法特征的汉语篇章分析平台，具体介绍了连接词识别和篇章语义关系分类这两个子任务的具体实现；第 5 节详细分析了实验结果；最后总结全文并指出下一步工作。

2 相关工作

随着 PDTB 的发布出现了很多英文篇章关系分析的相关研究。基于 PDTB 语料库的篇章分析工作主要包括连接词识别、论元标注、语义关系的分类以及隐式篇章关系识别等。其中显式篇章关系研究的代表性工作包括：

在连接词识别方面，Pitler 和 Nenkova（2009）使用最大熵模型，第一次将句法方面的特征（这些句法特征已经广泛应用于论元分类等任务中）应用到连接词识别任务中。在只有句法特征的情况下，连接词识别的 F_1 值达到了 88.19%。在此基础上，他们将连接词与句法特征相组合，获得了 94.19%的连接词识别 F_1 值。Lin 等（2012）在 P&N 的基础上新增了词法特征（包括连接词的词性和词与词性之间的组合）和两种句法路径作为特征。实验结果表明词法特征的加入进一步提高了英文连接词识别的性能，F_1 值达到了 95.36%。

在篇章关系识别方面，PDTB 将篇章关系分为四大类[5]。P&N 使用上述句法特征对英文显式关系的语义分类进行了研究，在 PDTB 上使用朴素贝叶斯分类器进行 10 倍交叉验证。实验结果表明只有连接词作为特征时，四类篇章关系的识别精度为 93.67%。加入句法特征后，识别精度提高到 94.15%。

与英文相比，其他语言或非新闻领域也有一些相关研究，典型工作包括：Alsaif 和 Markert（2011）[6]依照 PDTB 的标注框架对 APT（Arabic Penn Treebank）进行标注，并在此基础上研究阿拉伯语篇章中显式连接词的自动识别和篇章关系的分类。其中连接词识别的精度达到了 92.4%。Ramesh 等（2010）研究了在 PDTB 和生物语料库（BioDRB）上连接词识别的差异。他们使用条件随机场模型（CRFs）在 PDTB 上训练分类器，在 PDTB 和生物语料库上测试的 F_1 值分别为 84%和 55%。在生物语料上进行交叉验证的 F_1 值达到了 69%。

相比之下，对汉语显式篇章关系的研究相对较少，这主要是因为缺乏汉语篇章级别语料。我们依照 PDTB 框架[7]，选取 500 篇 CTB 文本进行了显式篇章关系的标注。汉语表达形式多样，篇章连接词的构成比英文复杂，这都给汉语显式关系分析造成了一定的困难。本文使用最大熵模型，结合词法、句法等特征，构建了汉语显式篇章关系分析平台，并通过实验分析了汉语篇章关系的复杂性。

3 汉语显式篇章关系语料库

目前可供研究的英文语料库主要有 RST Discourse Treebank（RST-DT）和 PDTB。RST-DT 由美国南加利福尼亚大学和华盛顿国防部联合标注，2002 年由 LDC（Linguistic Data Consortium）发布。它先利用 RST-Tool 工具对文本进行预标注，主要包括文本的切割（生成小句）和初始修辞关系的生成，然后人工验证预标注的结果，判断文本的切分是否正确，并为功能语句对标注一个可能性最大的修辞关系。

PDTB 是由 LDC 于 2008 年发布，是目前规模最大的英文篇章级别的语料库。PDTB 共标注了以下几种类型：（1）显式和隐式关系连接词；（2）Alternative Lexicalization（AltLex）；（3）Entity-based Coherence Relation（EntRel）；（4）No Relation（NoRel）。PDTB 还定义了一个三级层次的语义结构，第一层包括 Temporal、Contingency、Comparison 和 Expansion 四类语义，第二层包括 16 类语义，第三层包括 23 类语义。

与英文相比，汉语表达上更具多样性。参考 RST 理论，借鉴 PDTB 体系，我们选取汉语树库（Chinese Treebank，CTB）中的 500 篇新闻文本进行了汉语显式篇章关系的标注，共标注了 1690 个显式关系，标注内容主要包括连接词及其驱动的篇章关系的类别。

与英文 PDTB 体系类似，我们将汉语连接词也限定在某一范围内，设定了 258 个词构

成的连接词列表，并根据这些连接词在词语构成及语义表达上的主次关系选定了其对应的中心词，最终形成了 180 个连接词中心词列表。

在篇章关系方面，我们标注了四大类关系：因果类、并列类、转折类和解说类。每一类细分了具体的关系小类，共 17 个。汉语篇章关系的划分如表 1 所示。

语义类别	篇章关系
因果	因果关系、推断关系、假设关系、目的关系、条件关系、背景关系
并列	并列关系、顺承关系、递进关系、对比关系、选择关系
转折	转折关系、让步关系
解说	解说关系、总分关系、例证关系、评价关系

表 1 汉语篇章关系的划分

接着我们以汉语中出现频度较高的连接词“而”为例介绍汉语显式篇章关系的标注。表 2 给出的 6 个例子均摘自 CTB 语料，我们可以看到：句 1 中“而”作为连接词，表述的是转折关系；但在句 2 中“而”并不承担连接词角色。此外，“而”作为连接词，不仅可以表述转折关系，还可以表述其它语义关系。在例句 3-5 中的连接词“而”表述的语义关系分别为：递进关系、因果关系和例证关系。此外例句 6 中，“不……而”承担了篇章连接词的角色，但就连接词构成成份及其表述的语义关系看，“而”是这一连接词的中心词。

1. 外商投资企业的出口商品仍以轻纺产品为主，其中，出口额最大的是服装，去年为七十六点八亿美元。而进口商品则以机械设备和工业原材料为主。
（转折类.转折关系）
2. 为规范建筑行为，防止出现无序现象，新区管委会根据国家和上海市的有关规定，结合浦东开发实际，及时出台了一系列规范建设市场的文件，基本做到了每个环节都有明确而又具体的规定。
（不是连接词）
3. 推动经济增长的主要因素是亚洲地区经济发展依然强劲有力，全地区经济增长速度将达到百分之七点九，而中国增长速度可高达百分之九点七。
（并列类.递进关系）
4. 美国建议解除对波黑的武器禁运，意味着“国际援助可能结束”以及“重大的公开冲突再次爆发”，而这将产生“造成重大损失的严重后果”。
（因果类.因果关系；推断关系）
5. 每年仅能换回十多亿美元的外汇。而一九九六年，中国仅进口药物制剂就耗资超过十一亿美元。
（解说类.例证关系）
6. 一些外国金融机构在陆家嘴选址时，并不在乎楼盘的价格，而更注重大楼的档次与本公司的身价相符。
（并列类.选择关系）

表 2 连接词及其语义关系示例

4 汉语显式篇章关系分析平台

汉语表达形式多样，本节结合汉语特点给出了一个汉语显式篇章关系分析平台，由篇章连接词识别和篇章关系分类两部分构成。

4.1 篇章连接词识别

篇章连接词通常用来显式地表述基本文本单元之间承接的篇章关系。与英文类似，汉语中连接词候选也存在是否承担了篇章连接词角色的歧义。例如表 3 给出的示例，就连接词候

选“**和**”而言，例句 1 中的“**和**”承担了篇章连接词角色，表述的是并列关系，而例句 2 中的“**和**”并不是篇章连接词。所谓篇章连接词识别，正是针对连接词候选的这种歧义展开，主要任务就是确定连接词候选是否承担了篇章连接词的角色。

1. 此外还有一千多名幼稚园学生表演舞蹈庆祝香港回归祖国，**和**一百名演出者表演歌舞剧「释迦传」。
2. 而进口商品则以机械设备**和**工业原材料为主。

表 3 连接词候选是否承担篇章连接词角色的歧义示例

本文将连接词识别看作一个二元分类问题，首先根据语料标注中预定义的 180 个连接词的中心词列表获取连接词候选集，再针对每一连接词候选来选取特定的上下文特征，使用最大熵方法进行训练和预测[8]。

我们考虑的上下文特征主要包括词法和句法两方面。词法特征主要描述连接词及其所处的上下文词汇集的信息，而句法特征主要基于句法分析的结果获取连接词所在位置的句法信息[1]。此外我们还考虑了连接词与句法特征的组合以及多种句法特征间的组合信息。以表 2 中的句 1 为例，表 4 给出了连接词识别模块所使用的特征集，及其详细描述和取值情况。图 1 给出的是该例句中连接词候选“**而**”所处上下文的部分标准句法树。

类别	特征描述	取值
词法	词、词与前/后词的组合	“而”、“NONE_而”、“而_进口”
	词性、前/后词的词性及其组合	“AD”、NONE、“NN”、“NONE_AD”，“AD_NN”
句法	词的句法范畴	“ADVP”、“IP”、“NONE”、“NP-SBJ”
	词性节点到根节点的路径	AD_>_ADVP_>_IP
	词性节点到根节点的压缩路径	AD_>_ADVP_>_IP
组合	词与句法范畴的组合	“而_ADVP”、“而_IP”等
	句法范畴组合	“ADVP_IP”，“ADVP_NONE”等

表 4 特征集及其对应的描述

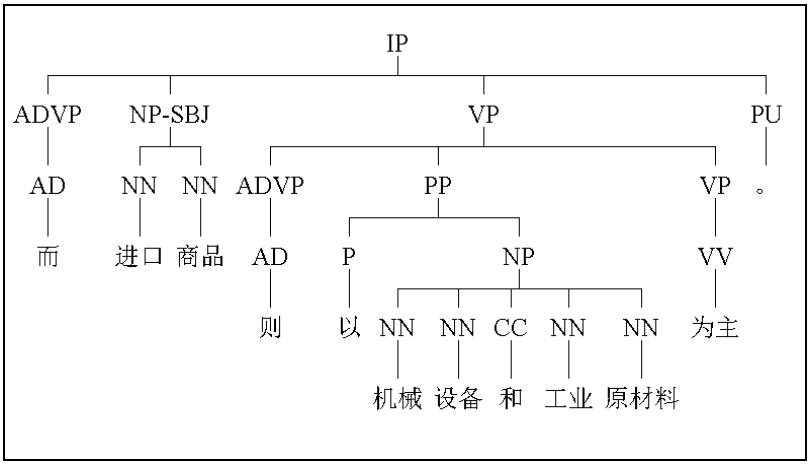

图 1 表 2 示例中例句 1 对应的标准句法树

4.2 篇章关系分类

通常连接词在篇章中表述某个特定的语义关系，例如“并且”承担连接词角色时，一般用来表述并列关系。但一些连接词在表述篇章关系时也存在语义上的歧义。例如表 5 中给出的两个例句，“**对此**”都承担了篇章连接词角色，但例句 1 中的“对此”表述的是因果关系，而例句 2 中的“对此”则表述了一个评价关系。篇章关系分类主要完成的工作是，根据连接

词及其所处的上下文判定其所表述的语义类别。

1. **对此**，国家税务总局副局长卢仁法在接受记者采访时强调，依法纳税是每个纳税人应尽的义务。

因果类.因果关系

2. 中泰两国的友好合作关系近年来继续顺利发展，两国在政治、科技、文化等各个领域的合作取得了显著成果。中国方面**对此**表示满意。

解说类.评价关系

表 5 篇章关系语义类别示例

本文将篇章关系识别看作一个多元分类问题。与连接词识别类似，我们针对每个连接词选定特定的上下文特征，再借助最大熵模型训练、判别连接词所表述的具体语义关系。由于显式关系是由篇章连接词驱动的，因此我们采用的特征包括词和它的词性。

5 实验与分析

构建了汉语显式篇章关系分析平台后，我们进行了实验验证。为了有效利用语料，我们采用 10 倍交叉验证的方法进行后续实验。实验中最大熵模型采用 OpenNLP 提供的 maxent 工具包[2]，参数均使用默认选项。标准句法树选自 CTB，自动句法树使用 Berkeley 句法分析器[3]获得（使用完全正确的分词）。评测指标方面我们采用标准的准确率（Precision）、召回率（Recall）和 F_1 值。

5.1 篇章连接词识别

本文采用了三类特征对连接词进行识别，表 6 给出了标准句法树下各类特征的贡献度。从表中可以看出只使用词法特征，连接词识别的 F_1 值达到了 65.9%。进一步考虑句法和组合特征，系统性能都有所提高。

特征组合	P（%）	R（%）	F1（%）
词法	71.92	60.81	65.90
句法	68.95	48.90	57.22
组合	71.84	56.19	63.06
词汇+组合	72.21	61.99	66.71
词汇+句法	73.05	61.46	66.75
词汇+句法+组合	72.73	61.75	66.79

表 6 标准句法树下各类特征的贡献度

表 7 给出了标准句法树和 berkeley 句法树下汉语篇章连接词识别的性能。我们可以发现篇章连接词识别的性能相差极小，即汉语篇章连接词识别性能对句法分析的性能好坏的依赖度较小[4]。

	P（%）	R（%）	F1（%）
标准句法树	72.73	61.75	66.79
berkeley	72.20	61.81	66.60

表 7 连接词识别的性能

与英文连接词识别的性能相比，汉语连接词识别性能比较低。为此我们对 180 个中心词

[2] http://maxent.sourceforge.net/

[3] http://code.google.com/p/berkeleyparser/

[4] 实际上这一结论与英文篇章连接词识别的研究一致。英文中自动句法分析对连接词识别 F_1 性能的影响小于 2%。

在语料中的分布情况进行了统计，其中有 76 个中心词在标注的显式关系中只出现了一次。我们对这 76 个中心词的识别情况进行了验证，发现由于训练实例较少，只有极少数被识别正确[5]。

5.2 篇章关系分类

由于显式篇章关系是由连接词驱动的，连接词在篇章关系语义类别的表述上起着关键性的作用。表 8 给出了语料库中标注的 1690 个显式关系在四大类上的分布情况，从中可以看到，并列关系比重最高，占到了一半以上，转折和解说类关系比重相对较低。

语义类别	标注的个数	所占比例（%）
因果	397	23.49
并列	850	50.30
转折	209	12.37
解说	234	13.84

表 8 语义类别的分布情况

我们采用词和词性作特征进行实验，表 9 给出汉语显式关系语义分类的性能。从实验结果可以看出，与英文显式关系类似，汉语显式关系的语义类别与连接词有很强的依赖，即确定了篇章连接词后，其语义一般没有歧义。

语义类别	P（%）	R（%）	F1（%）
因果	87.83	92.70	90.20
并列	93.21	95.29	94.24
转折	93.14	77.99	84.90
解说	93.81	90.60	92.17
合计	91.95	91.89	91.92

表 9 汉语显式关系语义分类的性能

所有的连接词中，部分连接词在标注的语料库中出现的次数相对较多。从表中我们可以看到，“并”、“其中”和“还”绝大多数实例都归为一类，构造出的分类器也确实将其归为了比重最高的一类。对于“而”，它的歧义最多，它在标注的语料中出现了 81 次，其中 57 次被标为并列类，约占 70.37%，分析实验结果发现，我们的分类系统将它也均归为了并列类，其分类性能的 F_1 值为 82.61%。

	在标注语料中出现次数	并列类	因果类	解说类	转折类
并	190	189	-	-	1
其中	155	3	-	152	-
而	81	57	2	1	21
还	68	67	1	-	-

表 10 出现频度较高（> 50 次）、有歧义的连接词的语义分布

6 结论

借鉴英文 PDTB 和 RST 语料，我们选取 500 篇 CTB 文本进行了汉语显式篇章关系的标注，基于这一语料构建了一个汉语显式篇章关系分析平台，并给出了实验结果及分析。为汉语篇章关系的分析奠定了良好的基础。但从实验结果看，与英文显式篇章关系分析相比，汉语篇章连接词识别的性能偏低。对此我们将尝试寻找新的符合汉语语言特性的特征，来提高汉语显式关系分析的性能。另一方面还将考虑利用汉语语料中标注的隐式关系，来辅助汉语

[5] 训练、测试集中去除这部分连接词候选后，系统 F_1 值提升了 0.5%，但不能通过显著性测试。

显式关系的分析。

参考文献

[1] PDTB-Group, 2007. The Penn Discourse Treebank 2.0 Annotation Manual[OL]. The PDTB Research Group.

[2] Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text[C]// In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Singapore.

[3] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2012. A PDTB-styled end-to-end discourse parser[J]. Natural Language Engineering.

[4] Ramesh Balaji. Hong Yu. 2010 Identifying discourse connectives in biomedical text[C]. //AMIA Ann Symp Proc 2010.

[5] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0[C]// In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

[6] Alsaif A, Markert K. Modelling discourse relations for Arabic[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 736-747.

[7] Xue N. Annotating discourse connectives in the chinese treebank[C]. //Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky. Association for Computational Linguistics, 2005: 84-91.

[8] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing[J]. Computational linguistics, 1996, 22(1): 39-71.