北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.029

汉语篇章连接词识别与分类

李艳翠[1,2] 孙静[1] 周国栋[1,†]

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;
† 通信作者, E-mail: gdzhou@suda.edu.cn

摘要 基于自建的汉语篇章结构语料库以及语料库中连接词及连接词关系类别的标注，抽取自动句法树和标准句法树的句法、词法、位置特征，利用有监督的方法进行连接词识别和分类。实验结果表明，连接词识别的 F1 值为 69.2%，连接词自动识别并分类的总正确率为 89.1%。

关键词 连接词识别；连接词分类；汉语篇章

中图分类号 TP391

Automatic Recognition and Classification on Chinese Discourse Connective

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

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

Abstract Based on the annotation of discourse connective in Chinese Discourse Treebank, especially the annotation of the connective and its relation classification. The authors extract syntax, lexical and position features of automatic syntax tree and standard syntax tree, using supervised method to recognize and classify connective. Experimental results show that connective recognition F1-measure is 69.2%, and connective classification accuracy is 89.1%.

Key words connective recognition; connective classification; Chinese discourse

自然语言的单位由小到大可以分为词、短语、句子和段落，最后形成篇章(Discourse)。篇章有时也称语篇或话语，指一系列连续的子句、句子或语段构成的语言整体单位。篇章不是语言成分的无序堆砌，每个篇章不仅具有内部连贯性，而且篇章中的各级单位是描述同一个问题或同一种情境的一个相对完整的语言整体。在一个篇章中，子句、句子或语段间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，才能对篇章有一个总体把握。篇章结构分析就是分析出篇章的层次结构及语义关系，它是自然语言处理的一个核心问题，也是近几年的研究热点和难点。篇章结构分析在自动文摘[1]、问答系统[2]、指代消解[3]、篇章连贯性评价[4]等方面都有所应用。

在汉语篇章中，篇章关系是指同一篇章内部，句子与句子或子句与子句之间的语义连接关系，如条件关系、转折关系、因果关系等[5]，连接词主要指连接不同单位并表示这种语义关系的词语。连接词在句子中一般不充当句法成分，没有修饰和限定作用，一般是表示连接作用的连词、关联词以及其他与之有同等关系作用的语言单位。本文所述的篇章连接词不限于现代汉语[6]中的连词，只要对句子

863 计划(2012AA011102)、国家自然科学基金(61273320)和河南省教育厅科学技术研究重点项目(14A520080)资助
收稿日期: 2014–06–29; 修回日期: 2014–10–28; 网络出版时间: 2014-12-01 09:26

和语段起连接作用，能恰当表示句子之间或子句之间关系的语言单位均可称为连接词。如例 1 中的“因此”、“对此”、“不是……而是”、“使”和“正因为”都是连接词，例句中字母引导基本篇章单位，“|”的个数表示基本篇章单位层次。

例 1 a 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，|| b 因此大量出现的是以前不曾遇到过的新情况、新问题。| c 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，||| d 而是借鉴发达国家和深圳等特区的经验教训，|||| e 聘请国内外有关专家学者，||||| f 积极、及时地制定和推出法规性文件，|||| g 使这些经济活动一出现就被纳入法制轨道。|| h 去年初浦东新区诞生的中国第一家医疗机构药品采购服务中心，正因为一开始就比较规范，||| i 运转至今，||||| j 成交药品一亿多元，|||| k 没有发现一例回扣。(chtb_0001)

从自然语言处理的角度分析篇章关系的前提是对连接词进行正确识别。基于显式篇章关系识别的目的，本文主要研究显式连接词的识别与关系分类。虽然有研究[7]表明连接词分类的效果较好，但都是在已知其是连接词的情况下进行分类的，本文采用自建的篇章结构语料库进行连接词识别与关系分类，探讨使用词法和句法特征进行连接词识别及分类，为篇章结构分析奠定基础。

1 相关研究

包含连接词标记的语料库目前主要有汉语复句语料库① (Corpus of Chinese Compound Sentences)、清华汉语树库(Tsinghua Chinese Treebank, TCT)[8]、哈工大中文篇章关系语料库② (HIT-CDTB)。汉语复句语料库由华中师范大学语言与语言教育研究中心开发，是一个面向汉语复句研究[9]的专用语料库。该语料库收有标复句 658447 句，约 44395000 字。语料来源以《人民日报》和《长江日报》为主，收入各种句式的现代汉语有标复句。在此语料上，胡金柱等[10]讨论利用关系词库中的信息来判断关系词的搭配关系、连用形式以及单用形式的方法，并从自建的 5000 条三句式复句语料集中抽取 1000 条进行关系词标注试验，正确标注的分句有 2073 个，分句的正确标注率达 69.1%。文献[11]结合词性标记和关系词搭配理论，提出正向选择算法提取关系词，测试结果表明，关系词提取的正确率达 89.8%。并非复句中出现的关系标记都是关系词，文献[12]利用汉语复句语料库和关系词库，提出一种基于规则的连用关系标记的自动标识算法，从中识别出真正的关系词，该算法结合关系词库和关系词提取技术，分析其连用特征，对连用关系标记标识正确率达 72.9%。

清华汉语树库[8]中标出了复句内各分句之间的关系信息，复句分类采用比较常用的分类方法，即并列关系、连贯关系、递进关系、选择关系、因果关系、目的关系、假设关系、条件关系和转折关系，但没有标注特定复句关系所对应的复句关系词。洪鹿平[13]在 TCT 上做汉语复句关系自动判断研究，穷尽式地收集关系词语，并把关系词语标注为联合和偏正两种类型，然后抽取特征，利用 CRF 模型进行分类。李艳翠等[14]利用句法及功能标注信息从 TCT 中提取复句关系词并标注其类别，然后抽取自动句法树和标准句法树的句法、词法、位置特征进行复句关系词的识别和分类，实验结果表明，复句关系词判断正确率达 95.7%，复句关系词类别判断 F1 值为 77.2%。

哈工大中文篇章关系语料[15]包括 525 篇标注文本，语料生文本来源于 OntoNotes 4.0，覆盖了句群关系、复句关系、分句关系等多级信息。标注采用宾州篇章树库的模式。标注的关联词分为显式关联词和隐式关联词两种。显式关联词包括普通关联词(“但是”、“由于”等)、带修饰关联词(“部分原因”、“尤其是”等)和平行关联词(“一方面……另一方面……”，“一边……一边……”等)。标注的篇章关系共 6 大类：时序关系、因果关系、条件关系、比较关系、扩展关系和并列关系。由于很多时间词(如“九八年”)被标注为连接词，所以语料中标注出

① http://ling.ccnu.edu.cn:8089/jiansuo/TestFuju.jsp

② http://ir.hit.edu.cn/hit-cdtb/index.html

来的显式关联词共有 1472 种, 它们总共出现的次数是 11519 次。在此语料上, 张牧宇等[16]进行句间语义关系的识别, 针对显式篇章句间关系, 提出基于关联词规则的方法进行识别, 取得了很好的效果。

Zhou 等[17]采用宾州树库模式标注了 98 篇汉语篇章语料, 所标语料只是初步尝试并没有正式发布。针对 Sinica Treebank3.1, Huang 等[18]采用 PDTB 的标注方法, 手工标注了 81 篇中文文章, 完成了 3081 个句对的小规模的中文篇章树库。但是, 他们以句子作为基本单位, 然而实际情况却更加复杂。Zhou 等[19]采用宾州篇章树库体系(PDTB)的方法标注了显式句内篇章连接词的论元和关系, 共标注 890 篇文档, 在标注时对篇章关系、论元范围、论元语义根据汉语特点进行了调整。

综上可知, 目前语料多关注句内关系, 所标连接词侧重句子内部, 忽略了句子之间的关系。目前的工作中连接词标注和分类多参考英语的分类方法, 与汉语分类体系相差较大, 标注出的连接词种类过多。本文主要采用自建的汉语篇章语料库进行汉语篇章连接词的识别与分类。

2 汉语篇章结构语料库

2.1 总体介绍

汉语篇章结构语料(Chinese Discourse Treebank, CDTB)采用树的形式表示汉语的篇章结构, 每一个段落构建一棵篇章结构树。例 1 的树形图如图 1 所示, 可以看到显式连接词“因此、对此、不是……而是、使、正因为”。CDTB 中标注了关系、连接词、中心、层次等信息[20], 关于隐式关系部分的标注可见文献[7]。本文主要介绍 CDTB 中连接词的标注情况。

例 1 连接词“因此”的标注信息如例 2 所示。其中, “<R”开头的每一行表示一个关系; “ConnectiveType”表示显式关系或隐式关系; “Connective”给出关系中的连接词; “RelationType”给出连接词的关系类型; “ConnectiveAttribute”表示显示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”表示包含的孩子节点。例 2 中的关系 ID 为“3”, 所表示的是“显式关系”, 连接词是“因此”, 属于“因果关系”。

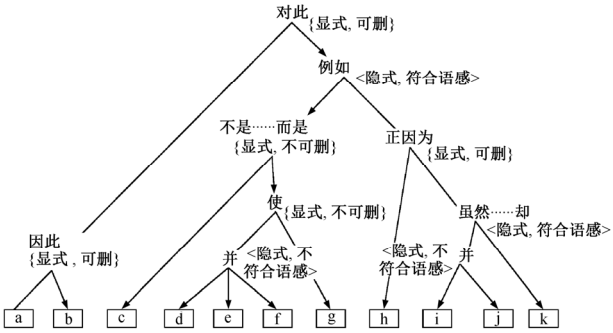

图 1 例 1 的篇章结构树

Fig. 1 Discourse structure tree of Example 1

例 2 <R ID="3" StructureType="逐层切分" ConnectiveType="显式关系" Layer="2" RelationNumber="单个关系" Connective="因此" RelationType ="因果关系" ConnectivePosition="37…38" ConnectiveAttribute="可删除" RoleLocation ="normal" LanguageSense="true" Sentence="浦东…工程, |因此大量…问题。" SentencePosition="1…36|37…60" Center="2" ChildList="" ParentId="1" UseTime="51"/>

CDTB 标注语料由两组标注者共同完成, 为便于一致性分析, 两组标注者对本语料的 chtb_0041~chtb_0100 共 60 篇文档分别进行标注, 根据标注结果计算得到: CDTB 子句分割标注一致性为 91.7%, Kappa 值[21]为 0.91; 显式关系和隐式关系判断一致性为 94.7%, Kappa 值为 0.81。显式连接词识别一致率为 82.3%。

标注初步完成后, 2014 年 6—8 月进行仔细校对, 目前已经标注了 CTB6.0 中的 500 个文档(chtb_0001~chtb_0657), 在 CTB 中句子标号从 1 到 6648。CDTB 目前共有效标注 2342 段, 每段平均 3 个句子。共有 10650 个子句(叶子节点), 平均每个有效标注的句子包含 2 个子句, 每个子句平均长度为 22 个汉字。CDTB 中共有 7310 个关系(二元或多元关系), 其中显式关系有 1812 个, 占 24.8%; 隐式关系有 5498 个, 占 75.2%。

2.2 连接词

连接词指具有子句及其以上语法单位连接和关系提示作用的语言单位。判断连接词的主要标准是看其连接的成分是否为子句(子句定义见文献[22])及其以上语法单位, 以及能否提示所连接篇章单位间的语义关系。作为篇章中的连接词, 主要作用是连接前后子句及前后篇章, 如例 1 中的“对此”连接

前后句子。作为连接词，在篇章中一般只起连接作用，不充当句法成分。如例 1 中的“因此”本身在句中对前后子句并不起修饰作用，只起到连接前后子句的作用。但也有既充当句法成分，也起连接作用的连接词。如例 1 中的“不是”，在子句中起否定作用，充当谓语成分，也对子句起连接作用。根据词的意义和所起的作用，对该类词做出判断也不难。此类连接词在连接词的总量中所占比例并不多。不像名词、动词、形容词具有一定的内涵，一般连接词没有实在的意义，如：“但是、不但……而且、因为……所以”等，起到连接前后子句的作用，词语本身并无实际意义。CDTB 中共标注显式连接词 274 个，出现次数最多的 10 个显式连接词及频率如表 1 所示。

语法性质上，连接词不限于传统连词，只要对子句及其以上语法单位的连接和关系有提示作用的语言单位均为连接词，连接词有连词、介词、副词等诸多语法类型[6]。连词起连接作用，连接词、分句、和句子等，表示并列、选择、递进、转折、条件、因果等关系，如“和、跟、同、与、及、或；而、而且、并、并且、或者；不但、不仅、虽然、然而、如果、因为、所以”等，在 CDTB 中，连词在所有连接词中所占的比例为 37.8%。副词修饰、限定动词、形容词性词语，表示程度、范围、时间等意义，如更、更加、还、还是、只、仅仅、尤其等。在汉语篇章结构中，利用副词的上述作用表示子句与子句间的递进、顺承等关系。如“尤其”作为连接词可以表示子句间的递进关系；“再”可以表示顺承关系等，副词在连接词中约占 26.5%。介词依附在实词或短语前面，共同构成介词短语，主要用于修饰、补充谓词性词语，介词常常充当逻辑成分，标明与动作、性状有关的原因、目的、方式、处所等，例如“因为、由于、为、为了、除了、对于、关于”，在 CDTB 中介词占 20.4%。

表 1　CDTB 中出现频率最多的显式连接词及次数
Table 1　Most frequent connectives in CDTB

连接词	次数	连接词	次数
并	208	其中	154
也	131	而	70
但	69	还	68
使	56	以	52
为	49	同时	46

在汉语篇章结构中，连接词的形式并不是单一的。根据形式不同可以分为独用连接词、关联词、合用连接词和其他类型。独用连接词是指子句与子句的连接词是单独的语词，如例 1 中的“因此”、“正因为”。独用连接词在连接词总数中占多数，CDTB 中独用连接词占 51%。关联词指子句与子句的连接词是成对出现的词语，如例 1 中的“不是……而是”，关联连接词占连接词总数的 35%。关联词也可以独立应用到子句中，单独表示子句间的连接关系，如“虽然……但是”也可以分别单独使用“虽然”和“但是”。合用连接词是指以几个词黏着的形式而组成短语，连接前后子句，表示篇章关系，如例 3 中的“从而也”是表示目的关系的“从而”和表示并列关系的“也”合用，再连接子句共同表示并列关系。“同时也、正因为、与此同时”都属于合用连接词，这类连接词占连接词总数的 2%。其他类型有“虽然……但……却、除……外……还、同时也、这表明”等形式，在 CDTB 占 12%。

例 3 他指出，美国国会每年就这个问题进行辩论实际上只有对美国自身不利，影响美国商人的对华投资信心，**从而也**影响到美国人的就业机会。

2.3　连接词的语义关系

连接词在连接篇章单元的同时，也表示它们之间的语义关系。CDTB 参考现代汉语[6]、复句研究[9]、汉语句群[23]的分类方法，将篇章关系分成 3 个意义层次：第 1 层 4 大类包括因果类、并列类、转折类和解说类；4 大类下面还包含第 2 层，共 17 个小类，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，具体关系分类层次见图 2。

从图 2 可知，4 大类中并列类有 972 个，所占比例最大，为 56.7%，最少的是转折类占 10.0%。17 个小类关系中，并列关系有 738 个，所占比例最大，为 40.7%，最少的是背景关系，仅有 4 个实例。

有些连接词属于不同的关系类型，如连接词“并”可以表示并列关系(CDTB 中共 199 个实例)、顺承关系(7 个实例)、解说关系(1 个实例)以及同时表示顺承和递进关系(1 个实例)。不同情况下可以表示不同关系的词共有 34 个，部分连接词及其可表示的关系如表 2 所示。在标注的过程中，连接词也会有同时认为既是甲关系又是乙关系的情况，如

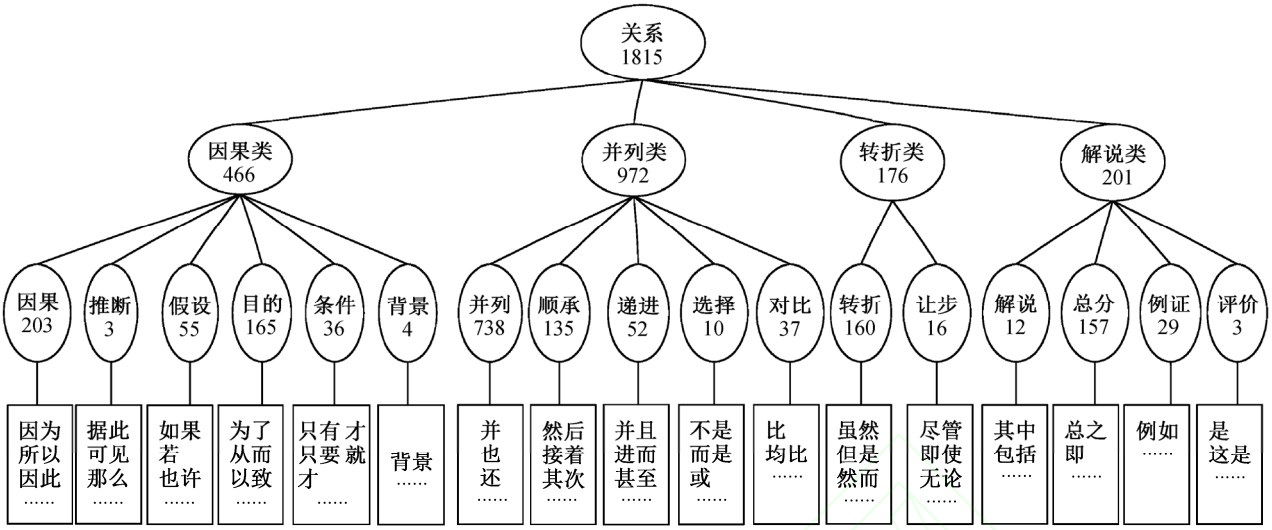

图 2　基于连接词的关系分类及分布

Fig. 2　Relation classification and distribution based on connectives

表 2　CDTB 中部分可表示多种关系的连接词及次数

Table 2　Connectives representing multi-relation in CDTB

连接词	表示关系(次数)
并	顺承关系; 递进关系(1)、顺承关系(7)、解说关系(1)、并列关系(199)
其中	解说关系(1)、总分关系(153)
也	顺承关系(2)、并列关系(128)、并列关系; 顺承关系(1)
而	递进关系(6)、顺承关系(1)、转折关系(5)、对比关系(18)、并列关系(39)、因果关系(1)
但	转折关系(66)、对比关系(3)
还	顺承关系(6)、并列关系(61)、选择关系(1)
使	因果关系(38)、目的关系(18)
如	例证关系(17)、假设关系(11)
又	顺承关系(10)、并列关系(16)
而且	递进关系(1)、并列关系(19)

“尽管……但”可以表示转折关系和让步关系, 并可表示顺承和递进关系, 也可表示并列和顺承关系, 这类实例较少, CDTB 中只标注了 3 个实例, 在统计时, 图 2 将每种关系分别统计为一个实例。

3　连接词识别与分类实验

李艳翠等[14]主要对清华汉语树库中的连接词进行识别与分类, 但清华汉语树库中并没有标注连接词信息, 而且清华汉语树库没有体现句子之间的连接词关系信息。参考文献[14]所用特征, 本文主要进行显式连接词的自动识别与分类实验, 所用 CDTB 以段落为单位进行标注, 不仅有连接子句的连接词, 还有连接句子的连接词。连接词在语料中有清晰的标注, 为连接词的识别与分类提供了资源。

实验抽取文献[14]提到的词汇、句法和位置特征, 所用标准句法树是 CTB6.0 语料中的句法树, 自动句法树是采用伯克利句法分析工具生成的句法树。使用 NLTK 工具包①进行实验, 使用 10 倍交叉验证方法识别连接词(二元分类)和连接词分类(多元分类)。

3.1　连接词识别实验与分析

对语料中标注的 274 个显式连接词, 抽取所有

① http://www.nltk.org/

出现这 274 个词的例子, 其中标注了连接词的为正例, 没有标注的为负例, 如“和”有时是篇章连接词, 为正例, 但大多数情况下为负例。对于联合连接词(如“不但……而且”), 将其处理为 2 个实例, 经处理后共有 226 个连接词, 如原来的“不但……而且”、“不但”、“而且”, 经处理后只剩“不但”和“而且”。实验共抽取 10524 个实例, 其中是连接词的实例有 2097 个, 非连接词实例为 8427 个。实验时将所有实例平均分成 10 份, 每次取 9 份训练, 剩下 1 份测试, 共进行 10 次。表 3 给出平均结果。

从表 3 可以看出, 使用决策树效果最好, 说明连接词识别问题并不太复杂, 可以抽取出一定的规则。单纯的词汇特征对连接词的识别也有一定作用。由实验可知, 利用词汇、句法和位置特征的组合进行连接词的识别效果最好, 使用自动句法树和标准句法树连接词识别的正确率分别为 88.4%和 88.5%。本文实验所用特征与文献[14]相同, 表 3 中给出文献[14]在清华树库上的部分实验结果。从表 3 可以发现, 使用决策树, 本文结果比文献[14]低 3.7%, 主要原因是文献[14]只考虑句内连接词, 连接词由算法抽取, 抽取时只考虑出现次数最多的连词(c)、副词(d)和连接词(l)三种词性, 本文语料采用自标的 CDTB, 连接词包括句内连接词和句间连接词。

表 3 的实验结果包含需要识别的词不是连接词的情况, 这类词所占比例较高(占 80%), 故而总正确率较高。表 4 给出利用决策树和最大熵分类器, 使用词汇、句法和位置特征对是连接词的词进行识别的准确率、召回率和 F1 值。从表 4 可知, 对于连接词的识别, 利用最大熵分类器的效果明显好很多。各种情况下对连接词的识别准确率均高于召回率。使用自动句法树的 F1 值为 69.2%, 使用标准句法树 F1 值 69.3%, 结果相差较小, 说明连接词识别任务对句法分析性能的好坏依赖程度较小。

表 4　连接词识别的准确率、召回率和 F1 值

Table 4　Connectives recognition Precision、Recall and F1-measure

分类器	类别	准确率/%	召回率/%	F1 值/%
最大熵	自动句法树	78.8	61.8	69.2
	标准句法树	78.9	61.8	69.3
决策树	自动句法树	56.8	49.6	52.3
	标准句法树	58.9	48.5	52.7

3.2　连接词分类实验与分析

连接词识别完成后, 需要对连接词进行分类, 本实验使用连接词识别时所用的特征, 利用最大熵, 采用 10 倍交叉验证, 分别进行给定连接词分类和自动识别连接词分类的实验。

给定连接词的分类实验共有 2097 个实例, 其中并列类 1165 个、因果类 464 个、转折类 256 个、解说类 212 个, 以所有实例均取并列类时为基准系统, 正确率为 55.5%。4 大类分类结果总正确率为 95.7%, 每种类别的识别结果如表 5 所示。

从表 5 可以发现, 所有类别结果均远远好于基准系统。解说类、并列类识别效果较好, 因为解说类有比较明显的连接词(如“例如”), 并列类所占比

表 5　给定连接词 4 大类别识别结果

Table 5　4 Categories of given connective classification results

类别	准确率/%	召回率/%	F1 值/%
因果类	83.8	68.4	75.1
转折类	78.5	59.6	67.0
并列类	82.5	93.6	87.7
解说类	89.7	82.8	85.9

表 3　是否为连接词识别正确率

Table 3　Connectives recognition accuracy　　单位: %

语料库	特征	自动句法树			标准句法树		
		最大熵	决策树	贝叶斯	最大熵	决策树	贝叶斯
本文 CDTB	词汇	86.2	87.3	81.5	86.7	87.2	81.8
	句法	80.9	82.2	80.3	84.3	84.7	82.4
	词汇+句法	86.6	88.1	81.9	87.9	88.9	83.7
	词汇+句法+位置	87.2	88.4	83.4	88.2	88.5	85.3
清华树库[14]	词汇+句法+位置	91.2	92.1	88.1	–	–	–

例较大, 识别效果也较好。转折类识别效果最差, 部分原因是一些转折类的词也可以表示并列关系, 例如“而”既可表示转折, 又可表示并列, 并列关系所占比例较大(见表 2), 影响结果的判断。

通常, 我们对连接词分类是在并不知道其是否为连接词的情况下进行, 因此首先需要确定某个词是否为连接词, 然后对识别为连接词的词进行分类。测试时首先使用连接词识别分类器识别实例是否为连接词, 若是则使用连接词分类器给出类别, 实验同样进行 10 次取平均值, 得到连接词分类总正确率为 89.1%, 明显低于给定连接词的总正确率 95.7%。表 6 给出在连接词自动识别基础上进行连接词分类的准确率、召回率和 F1 值。表 6 是在连接词自动识别正确的情况下进行连接词分类的结果, 识别出连接词后对其判断类别相对容易。

分析实验结果, 我们发现大部分篇章连接词只表示一种关系, 有 34 个篇章连接词(占所有连接词的 12.4%)可以表示多种关系, 识别错误主要由这类连接词歧义导致。这类连接词总数虽然不多, 但常用连接词较多(如“并、而、但”)。同一连接词对应的关系类型越少、类型越集中, 该词的歧义性越小。大部分连接词的歧义性较小, 80%以上指示同一种关系大类。以连接词“而”为例, 它对应的具体关系类型及关系大类分别如下: 并列类 64 次(其中并列关系 39 次, 对比关系 18 次, 递进关系 6 次, 顺承关系 1 次); 转折类 5 次(转折关系 5 次); 因果类 1 次(因果关系 1 次)。因此, 连接词识别与分类的主要难点是连接词判断的歧义及连接词关系类别分类的歧义。

表 6　自动连接词识别及 4 大类别识别结果

Table 6　4 Categories of connective recognition and classification results

类别	准确率/%	召回率/%	F1 值/%
因果类	72.8	80.5	76.2
转折类	73.2	70.8	71.2
并列类	64.7	95.8	77.2
解说类	82.5	86.7	84.5

4　总结

本文主要进行汉语篇章连接词的识别与分类。实验采用自建的汉语篇章结构语料, 语料中包含连接词及其类别的标注信息。根据标注结果, 识别抽取出的词是否为连接词, 并为其标示类别, 形成实验数据。数据抽取完毕后, 分别使用自动句法树和标准句法树中的句法、词性、位置特征进行连接词的识别与分类。实验表明, 连接词识别对句法分析性能依赖不大, 使用自动句法树, 关系词识别 F1 值为 69.2%。如果给定连接词, 连接词分类的总正确率为 95.7%; 在连接词自动识别的情况下, 连接词分类的总正确率 89.1%, 说明连接词识别性能对连接词分类结果有较大影响。目前的工作仍有需要改进的地方, 下一步准备进一步完善工作, 更准确的识别连接词及其类别。

参考文献

[1] 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. Portland: Association for Computational Linguistics, 2010: 147–156

[2] Verberne S, Boves L, et al. Discourse-based answering of why-questions. Traitement Automatique des Langues, Discours et document: treatments automatous, 2007, 47(2): 21–41

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

[4] 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-Volume 1. Portland: Association for Computational Linguistics, 2011: 997–1006

[5] 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, 2013, 27(2): 17–25

[6] 黄伯荣, 廖序东. 现代汉语(下册). 北京: 高等教育出版社, 2002

[7] 孙静, 李艳翠, 周国栋, 等. 汉语隐式篇章关系识别. 北京大学学报: 自然科学版, 2014, 50(1): 112–117

[8] 周强. 汉语句法树库标注体系. 中文信息学报, 2004, 18(4): 1–8

[9] 邢福义. 汉语复句研究. 北京: 商务印书馆, 2001: 1–37

[10] 胡金柱, 吴锋文, 李琼, 等. 汉语复句关系词库的建设及其利用. 语言科学, 2010, 9(2): 133–142

[11] 胡金柱, 舒江波, 姚双云, 等. 面向中文信息处理的复句关系词提取算法研究. 计算机工程与科学, 2009, 31(10): 90–93

[12] 胡金柱, 陈江曼, 杨进才, 等. 基于规则的连用关系标记的自动标识研究. 计算机科学, 2012, 39(7): 190–194

[13] 洪鹿平. 汉语复句关系自动判断研究[D]. 南京: 南京师范大学, 2008

[14] 李艳翠, 孙静, 周国栋, 等. 基于清华汉语树库的复句关系词识别与分类研究. 北京大学学报: 自然科学版, 2014, 50(1): 118–124

[15] 张牧宇, 秦兵, 刘挺. 中文篇章级关系体系及类型标注. 中文信息学报, 2014, 28(2): 28–36

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

[17] Zhou Y P, Xue N W. PDTB-style discourse annotation of Chinese text // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Jeju: Association for Computational Linguistics, 2012: 69–77

[18] Huang H H, Chen H H. Chinese discourse relation recognition // Proceedings of the 5th International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing. Chiang Mai, 2011: 1442–1446

[19] Zhou L J, Li B Y, Wei Z Y, et al. The CUHK Discourse TreeBank for Chinese: annotating explicit discourse connectives for the Chinese TreeBank // Proceedings of the International Conference on Language Resources and Evaluation. Reykjavik, 2014: 942–949

[20] Li Y C, Feng W H, Kong F, et al. Building Chinese discourse corpus with connective-driven dependency tree structure // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, 2014: 2105–2114

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

[22] 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

[23] 吴为章, 田小琳. 句群. 上海: 上海教育出版社, 1984