英语语篇结构分析研究综述*

李艳翠[1,2], 朱坤华[2], 周国栋[1]
(1. 苏州大学 计算机科学与技术学院, 江苏 苏州 215006; 2. 河南科技学院 信息工程学院, 河南 新乡 453003)

摘 要: 全面系统地分析了英语语篇结构分析的相关理论、语料资源及国内外的相关研究成果,给出了语篇结构分析的研究趋势,为英语和汉语语篇结构分析研究做了基础性的工作。

关键词: 语篇结构分析; 语篇关系; 修辞结构理论; 宾州语篇树库
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)06-2018-06
doi:10.3969/j.issn.1001-3695.2012.06.004

Summary of research on English discourse parsing

LI Yan-cui[1,2], ZHU Kun-hua[2], ZHOU Guo-dong[1]
(1. *School of Computer Science & Technology, Soochow University, Suzhou Jiangsu* 215006, *China*; 2. *School of Information Engineering, Henan Institute of Science & Technology, Xinxiang Henan* 453003, *China*)

Abstract: This paper interpreted the concepts of English discourse parsing, introduced the corpus resource and gave the domestic and international research results. Finally, it discussed the future study tendency. This is a fundamental work for English and Chinese discourse parsing research.

Key words: discourse parsing; discourse relation; rhetorical structure theory; Penn discourse treebank

0 引言

关于自然语言的研究,很长时间人们把目光放在字、词、句子的层面上,并取得了显著的成果。随着研究的深入,需要考虑句子和句子的关系,则进入到语篇单位层级,看句子和句子如何组成更大的语言单位——语篇文本。在语篇这个层面上,要回答句子和句子如何组合,就是要解决篇章衔接和连贯的问题,即如何联句成篇的问题。语篇(discourse)指的是实际使用的语言单位,是交流过程中的一系列连续的语段或句子所构成的语言整体,也称为篇章或话语。总的说来,语篇由一个以上的语段或句子组成,其中各成分之间,在形式上是衔接(cohension)的、在语义上是连贯(coherence)的,衔接和连贯是语篇的基本要素和重要特征。语篇接受者对其理解的透彻度在很大程度上取决于对语篇连贯性的感知度,研究语篇中句子排列、衔接和连贯,是一种超句法分析。语篇结构是句子之间的有结构的表示,语篇结构分析是一个比较复杂的问题,分为基本语篇单位识别、语篇关系识别和语篇结构生成三部分,目前的研究主要集中在基本语篇单位识别和语篇关系识别上。人们普遍认为,在一篇文章中,句子或从句只有通过一定的关系相互连接才能更好地体现文章的主要内容,只有做好语篇结构分析才能更好地让计算机理解自然语言。自动语篇结构分析对很多自然语言处理任务都非常重要,如自动文摘、指代消解、情感分析、问答系统和对话生成等。

1 相关理论

关于英语语篇理论的研究比较多,本文所述主要是针对语法结构较强的书面语。

1.1 浅层的衔接关系

在语言学研究中,Halliday 等人[1]最早将衔接的各种修辞手段作为一种专门的语言现象进行系统的分析,并对此类衔接手段进行详尽的研究。他们合著的《英语的衔接》标志着衔接理论的创立,衔接理论的建立是语言学理论创建的有益探索。衔接是一个语义概念,它指的是语篇中语言成分之间的语义联系,或者说是语篇中一个成分与另一个可以与之相互解释的成分之间的关系。当语篇中一个成分的含义依赖于另一个成分的解释时,便产生衔接关系。作者提出了衔接手段的分类:a)语法衔接(grammatical cohesion),包括照应(reference)、替代(substitution)、省略(ellipsis)和连接(conjunction);b)词汇衔接(lexical cohesion),包括重复(repetition)、同义/反义(synonymy/antonymy)、上下义/局部—整体关系(hyponymy/meronymy)和搭配(collocation)。其中连接是运用连接成分体现语篇不同成分之间具有何种逻辑关系的手段,将句子作为语篇连接的基本单位,初步将连接成分划分为四种类型,即加合(additive)、转折(contrastive)、因果(causal)和时间(temporal)。后来 Halliday 等人对语篇中连接成分的分类采用了新的分类方法,即分为详述(elaboration)、延伸(extension)和增强(enhancement)。

1.2 Hobbs 的连贯关系

Hobbs 模型[2,3]提出,语篇结构由语篇单位 discourse unit 和语篇连接关系(coherence relations)构成。语篇单位可以小到子句,大到语篇本身。语篇关系表示两个语篇单位之间的语义关联性,有 12 种关系类型,部分语篇关系定义为(设 S0 和 S1 为两个相关的句子):a)结果关系(result),推测 S0 所声明的状态或事件(可能)导致 S1 所声明的状态或事件;b)解

收稿日期: 2011-10-24; **修回日期**: 2011-12-27 **基金项目**: 国家自然科学基金资助项目(90920004,60873150,60970056)

作者简介: 李艳翠(1982-),女,河南新乡人,助教,博士研究生,主要研究方向为自然语言处理(yancuili@gmail.com);朱坤华(1974-),女,副教授,硕士,主要研究方向为智能信息处理;周国栋(1967-),男,教授,博导,主要研究方向为自然语言处理、多语言跨文本信息抽取、网络信息挖掘.

释关系(explanation),推测 S1 所声明的状态或事件(可能)导致 S0 所声明的状态或事件;c)并列关系(parallel),推测 S0 所声明的 $P(a_1,a_2,\cdots)$ 与 S1 所声明的 $P(b_1,b_2,\cdots)$ 是类似的;d)细化关系(elaboration),推测 S1 和 S0 所声明的是同一命题 P;e)时机关系(occasion),推测由 S0 所声明的状态到 S1 最终状态的变化,或者由 S1 所声明的状态到 S0 的最初状态的变化。

Hobbs 提出的关系和它们的语义表示对其他研究者有较大的影响,包括 discourse graphbank、SDRT 和 Penn discourse treebank 都借鉴了 Hobbs 模型。

1.3　修辞结构理论

修辞结构理论(rhetorical structure theory, RST)是美国学者 Mann 等人[4,5]在系统功能理论的框架下创立的篇章生成和分析的理论。RST 与 Hobbs 模型有很多的相似性,其定义了 4 大类 25 小类语义关系,称为修辞关系。每个修辞关系的定义包括限制条件和效果两个部分。修辞关系具有相对开放性的特点,这些关系是 RST 的核心,也是语篇连贯的重要标志。

每个修辞关系可以连接两个或多个篇章单位。通常修辞关系连接的单位存在主次之别,其中表示主要信息的单位称做核(nucleus),表达次要信息的单位称做卫星(satellite)。这类关系也称为"单核"修辞关系。也有一些修辞关系连接的单位中无主次之分,如对比关系(contrast)和列表关系(list)。这类关系称为"多核"关系。如例 1 中有两个语篇单位(1A 和 1B),其中 1A 是核心单位,在图 1 中用竖线标志,1B 是卫星单位,二者之间是非意愿性原因关系。

例 1　[Mary is in a bad mood.]^{1A} [because her son is ill.]^{1B}。

在 RST 中,当两个以上的语篇单位形成修辞关系时,就构成了一种树结构——修辞结构树。句子与句子之间构成一种关系,从而形成一个大的单位,与相邻的单位再构成更高层的修辞关系,继而得到所谓的层次化语篇结构树。每个语篇的层次多少是不固定的,层次的多少是由语篇中句与句之间的语义关系的复杂程度决定的。通常情况下,语义关系越复杂,层次就越多。由于修辞关系含有特定的语义,语篇结构关系也就表达了语篇内部的语义关系。相比 Hobbs 模型,RST 更注重句子内部的篇章结构,篇章单位可以小到短语,在篇章计算方面受到了较多的关注。

1.4　基于图的方法

Wolf 等人[6]提出了将语篇表示为一个链图,图中的弧可以是有向的也可以是无向的,其语义关系大致遵循 Hobbs 的语义关系。由两个标注者标注了 135 篇新闻文章,标注这些关系没有任何关于如何标注的先见知识,给出一组如何分割文本、标注知识的指南。标注者按照如下步骤执行:a)将文本分割成基本的语篇单位;b)当语篇单位承当关系的论元时将语篇单位分组为更大的单位;c)识别两个片段之间是否有语篇关系;d)识别语篇关系的具体类别。

图 2 是将例 2 表示为图的一个实例。

例 2　a. Susan wanted to buy some tomatoes
　　b. and she also tried to find some basil
　　c. because her recipe asked for these ingredients.
　　d. The basil would probably be quite expensive at this time of year.

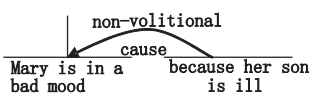

图1　包含核心结构的1A和辅助结构的1B的因果关系

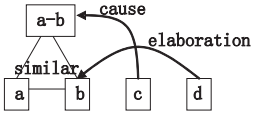

图2　一个图的实例

有很大一部分图库中的交叉关系需要建立实体级别的衔接。许多交叉是由于归因关系,这虽然是语篇关系,但与 Hobbs 定义的关系有本质的区别。然而,用图结构表示语篇也引起关于语篇表示的讨论,用树结构来表示语篇有一定的局限性,因为交叉关系在实际文本中是普遍存在的。

2　语料资源

2.1　修辞结构理论语篇树库

英语 RST 语篇树库[7]以 RST 为支撑,构建语篇标注语料库 RST discourse treebank。所标注的 385 篇华尔街日报文章皆取自宾州树库,篇幅长度不等,从 31 个词到 2 124 个词,总词数达到 176 000,平均每篇文章 458 个词。文章的内容涉及各种话题,如财政报道、商业新闻、文化点评、编者按、读者来信等。

语篇结构分析的第一个任务是确定基本语篇单位(elementary discourse units,EDU)。在具体标注时,研究者对基本语篇单位的识别规定为:充当主语或宾语的从句不属于基本语篇单位;充当主要动词补语的从句不属于基本语篇单位;所有词汇或句法标记的起状语作用的从句都属于基本语篇单位,包括起状语作用的非谓语动词词组;定语从句、后置的名词修饰短语或将其他基本语篇单位割裂开的从句或非谓语动词短语为内置语篇单位;有明显语篇标记的短语作为基本语篇单位,如因为(because)、尽管(in spite of)、根据(according to)等引导的短语属于基本语篇单位。

基本语篇单位确定下来后,剩下的工作就是根据 RST 确定基本语篇单位之间的关系,进而生成有层次的语篇结构树。语篇关系可能是单核关系或多核关系。文献[5]提出 RST 时只给出了 20 多种修辞关系,但他们明确指出这是一个开放关系集,既然是开放性的,就意味着读者在给定话语的内部可以定义出其他的关系类型。RST 语篇树库中标注了 53 种单层核心关系和 25 种多层核心关系,78 种关系又分成 16 个组别,每组都具有相同的修辞功能,如原因、比较、对比、条件等。

为了建立高质量的前后一致的标注标准和方法,研究者采用人工标注的方法,所选的标注者都是有过标注经历的、从事语篇分析和新闻报道的专业人员。在正式标注之前,他们都接受了专门的语篇结构标注培训。在整个语料库的建设过程中,研究者一直设法保证标注者之间的内部一致性。

2.2　宾州语篇树库

2.2.1　介绍

宾州语篇树库(Penn discourse treebank,PDTB)是在宾州树库和 PropBank 的基础上发展起来的,目标是开发一个标注有语篇结构信息的大规模语料库,主要标注与语篇连接词(discourse connectives)相关的连贯性关系(coherence relation)。标注信息主要包括连接词的论元结构、语义区分信息以及连接词和论元的属性相关特征等。

PDTB 是一个在宾夕法尼亚大学研究的美国自然基金资助项目,这个项目的目的是标注 100 万字的《华尔街日报》Treebank-2 语料库(LDC95T7)中所提供的语篇关系,目前版本

是 PDTB 2.0[8]。PDTB 2.0 标注了以下几种论元结构关系:显式关系(explicit)、隐式关系(implicit)、替代关系(altLex)、实体关系(entRel)、无关系(noRel)。

除了论元结构关系,PDTB 2.0 为每种论元结构关系标注了语义区分信息,同时也为每个语篇关系捕捉多义的连接,为每一种关系和它们的论元标注属性信息。显式关系、隐式关系、替代关系都标有语义区分信息和属性信息,实体关系和无关系无标注。PDTB 是一个标注了语篇结构和语篇语义相关信息的大规模语料库。虽然可以从话语的许多方面来完整地理解自然语言,PDTB 专注于编码语篇关系。PDTB 的标注方法遵从词汇化基础方法,采用与理论无关的做法,目的是为了让语料在不同的理论框架内使用。

PDTB 基于《华尔街日报》,共标注了 2 304 篇文章,约 100 万词,标注的文章主要以新闻为主。PDTB 2.0 总共有 25 部分,建议使用 2~21 部分作为训练集,22 部分作为开发集,23 部分作为测试集;0、1、24 如果需要的话可以作为附加的开发集。

2.2.2 话语关系及论元标注

PDTB 2.0 在 PDTB 1.0 的基础上对标注进行了扩张和修订,主要是对连接词的语义信息、连接词及其论元的属性信息方面进行了扩展,目前的研究大都在 PDTB 2.0 上进行。

因为语篇连接词的论元没有像动词那样公认的抽象语义分类的参数(实施者、受施者、内容),对语篇连接关系的两个论元,简单地记为 arg1 和 arg2。其中,arg2 是指出现在从句中和连接词在句法上相邻的论元,arg1 是另外一个论元。下面的例子中,arg1 用斜体表示,arg2 用黑体表示,连接词用下画线标棱。PDTB 中的话语连接关系包括:

a)显式语篇连接关系(explicit discourse connectives),句子中有明确的连接词标志,例如:*The city's Campaign Finance Board has refused to pay Mr. Dinkins $95,142 in matching funds* because **his campaign records are incomplete.**(0041)。

b)隐式语篇连接关系(implicit discourse connectives),在句子之间没有明确的连接词,但语篇关系是可以推断出来的,可以在句子中插入一个连接词,句子仍然通顺。例如:*Motorola is fighting back against junk mail. So much of the stuff poured into its Austin, Texas, offices that its mail rooms there simply stopped delivering it.* (so) **Now, thousands of mailers, catalogs and sales pitches go straight into the trash.**(0989)。

c)替代关系(altLex),语篇关系是可以推断的,但是插入连接词会显得冗余。例如:*After trading at an average discount of more than 20% in late 1987 and part of last year, country funds currently trade at an average premium of 6%.* AltLex[The reason:] **Share prices of many of these funds this year have climbed much more sharply than the foreign stocks they hold.**(0034)。

d)实体关系(entRel),两个句子没有语篇关系,第二句话只用为了对第一句话中的某个实体提供进一步的描述。例如:*Pierre Vinken will join the board as a nonexecutive director Nov. 29.* EntRel **Mr. Vinken is chairman of Elsevier N. V., the Dutch publishing group.**(0001)。

e)无关系(noRel),没有语篇关系和实体一致性可以从相邻的句子中推导出来。例如:*Mr. Rapanelli met in August with U. S. Assistant Treasury Secretary David Mulford.* NoRel **Argentine negotiator Carlos Carballo was in Washington and New York this week to meet with banks.**(0021)]。

PDTB 2.0 中共标注了 40 600 种关系,表 1 中给出了每种关系的具体分布情况。

表 1 PDTB2.0 关系数量分布

PDTB 关系	数量	PDTB 关系	数量
explicit	18 459	entRel	5 210
implicit	16 224	noRel	254
altLex	624		

2.2.3 语义标注

PDTB 对显式连接关系、隐式连接关系、替代关系都标注了语义类别信息。与动词一样,语篇连接词同样有不止一个语义,具体语义依赖于上下文和论元的内容。例如,since 有三种语义,一个是纯粹的时间,另一个是纯粹的原因,还有一个是原因和时间都是。

a) *The Mountain View, Calif., company has been receiving it was de-mon-*1,000 *calls aday about the product* since **strated at a computer publishing conference several weeks ago.**

b) *It was a far safer deal for lenders* since **NWA had a healthier cash flow and more collateral on hand.**

c) *Domestic car sales have plunged* 19% since **the Big Three ended many of their programs Sept. 30.**

在这种情况下,语义标注的目标主要是区分连接词到底是哪种语义。在所有情况下,语义标注指出了连接词所连接的论元之间的关系,当两个论元有两个以上的关系时,多种语义关系被同时标注。

PDTB 语料库中所标注的语义信息采用层次化的方法来组织,分别是类别、类型、子类。最顶层的类别代表主要的语义类别,分为四类,即 temporal、contingency、comparison 和 expansion;对每一个类别,一系列的类型被定义出来细化这个类别,例如,temporal 有 asynchronous 和 synchronous;第三层的子类用来指定每一个论元的语义贡献。图 3 是 PDTB 的语义标注层次示意图,共 4 大类别、16 个类型、23 个子类。

2.3 Discourse GraphBank

文本有时并不能像 RST 假设的那样表示为树型结构,如交叉依赖和节点有多个父节点的情况在文本中也比较常见,这种情况就不能构建树型结构。为了克服这个问题,Wolf 等人提出用图来表示文本,这种表示方法与实际情况更接近。他们标注了 135 篇文档,其中 105 篇来自 AP Newswire,30 篇来自 WSJ,形成了语篇图库。他们将从句标注为基本的语篇单位,基本语篇单位在语篇图中表示为一个节点。语篇图库中的修辞关系共 11 种:cause-effect,condition,violated,expectation,elaboration, example, generalization, attribution, temporal sequence, similarity,contrast and same。如图 2 所示,两个节点的边表示一个关系,关系可以是有向的非对称关系,如原因和条件,也可以是无向的对称关系,如并列和对比。

3 国内外研究现状

3.1 在 RSTDT 上的研究现状

在 RST 语篇树库上进行语篇结构分析研究一般分为基本语篇单位识别和语篇结构生成两步。

3.1.1 基本语篇单位识别

关于 EDU 的自动识别研究者研究得较多,结果也比较理想,Soricut 等人[9]采用基于统计的方法进行识别,识别器由两部分组成,第一部分是一个统计模型 $P(b_i|w_i,T)$,其中 T 是句子的句法树,w_i 是句子中的词,b_i 表示给定词是否是边界。如果给定的词是边界则 $P=1$,句子边界也是 EDU 边界。作者在

计算概率时用到了词汇和句法特征。第二部分是分割模块,分割 EDU 的原则是上述词边界的概率值大于 0.5。基本语篇单位识别在自动句法树上获得的 F 值为 83.1%,在标准句法树上的 F 值为 84.7%。由于文献[9]的方法不包含线索词,所以不能准确地识别复杂句子的边界。Huong 等人[10]提出了一个采用句法和线索词进行 EDU 分割的方法,给出了一个基于规则的基本语篇单位识别器,分割采用多步算法,使用句法信息和语篇线索词,得到的 F 值为 80%,EDU 识别不用任何训练。但该方法只用了 8 篇文章进行测试,结果不具有代表性。Subba 等人[11]使用神经网络进行 EDU 识别,在自动句法树上的 F 值 84.41%,在正确句法树上的 F 值为 86.07%。Milan 等人[12]基于句法和词法信息进行语篇分割,认为使用规则进行语篇分割比使用自动学习的方法有某些优势。实际上,作者给出的语篇分割模型不依赖于特定的训练语料,但获得了较高的准确度,主要因为插入了少量但高质量的边界信息。此系统中并没有按照 RSTDT 中语篇分割的方法进行分割,避免了构建短基本语篇单位,因其包含信息较少;补语从句不作为 EDU,如"He said that"不认为是 EDU。Hernault 等人[13]给出了一个基于序列数据标注的语篇分割模型,使用词汇和句法特征,采用 CRF 训练 RSTDT。系统与其他基于规则、基于统计以及基于 SVM 的语篇分割模型相比有一定的优越性,实验表明该序列模型语篇分割 F 值结果是 94%,接近于人工语篇分割的 F 值 98%。由上可知,RSTDT 基本语篇单位识别目前准确率较高,进一步提升的空间不大。

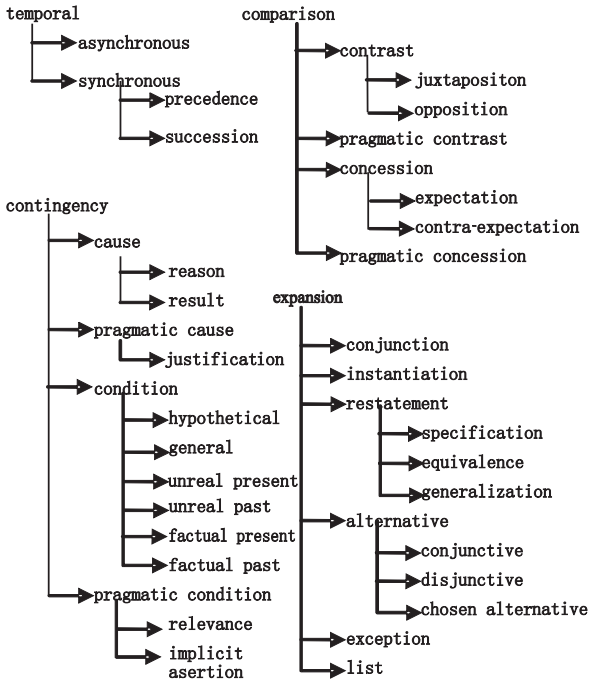

图3　语义标注及其层次

3.1.2　语篇结构生成

语篇结构生成是语篇结构分析的重点,也是难点。Soricut 等人[9]利用语法和词法信息进行句子级的语篇结构分析,其算法称为 SPADE,在语篇关系识别时采用一个概率模型计算各种语篇关系的概率。语篇结构分析模型采用全自动的方法,识别无标注的语篇关系的准确率是 70.5%,采用正确的基本语篇单位和正确句法树的结果是 96.2%。但 SPADE 并不对整篇文本进行语篇关系识别,通过实验证明句法和语篇信息的关系,在识别时并没有用到线索词。Huong 等人[14]研究了书面文本语篇结构的生成,给出了一个书面文本自动语篇结构生成系统,系统分为两个层次:句子级的语篇结构分析和文本级的语篇结构分析。句子级的语篇结构分析使用句法和线索词来进行语篇基本单位识别和语篇结构生成。对于篇章级别,为降低篇章结构分析的搜索空间,加入了文本相邻和文本组织限制。实验采用 RSTDT,结果表明降低搜索空间后 F 值为 70.1%,但计算量较大。DuVerle 等人[15]提出了一个基于支持向量机的语篇结构分析器,给出了一种新的方法进行语篇结构分析,采用统计机器学习的方法,使用了丰富的特征空间,给出的篇章结构树不仅仅是句子级的。该方法并没有进行 EDU 识别,而是采用现成的工具进行 EDU 识别,主要原因是作者认为基本语篇单位的识别准确率较高,所以只关心第二步的工作。采用两个分类器,其中 S 是一个二元分类器,L 是一个多元分类器。如果用 S 判断出相邻的两个 EDU 之间存在关系,则用 L 进行修辞关系识别和核心部分标注,关系识别的 F 值为 73.9%。该文献同时给出了一个构建全局结构树的方法,但到底采用怎样的特征空间需要进一步探索。

3.2　在 PDTB 上的研究现状

在 PDTB 语料上的研究一般分为三步:a)识别存在语篇连接关系的句子,这些句子可能有连接词,也可能没有连接词;b)识别存在语篇连接关系句子的论元 arg1 和 arg2;c)确定连接词所连接的两个论元之间具体的语篇关系。目前关于是否存在语篇关系的研究较少,一般是采用语料库中标注好的存在语篇关系的句子进行实验。因为识别语篇关系的论元 arg1 和 arg2 比较复杂,所以研究结果不太理想,目前研究主要集中在第三步。

3.2.1　论元识别

文献[16]自动识别语篇中的论元,将识别 PDTB 中标注的存在语篇关系的全部论元简化为只识别论元的主要部分,即中心词信息,采用句法树、依存树、连接词、词汇等多种特征进行论元识别,在人工标注的句法树上获得了 74.2% 的准确率,在自动获得的句法树上获得了 64.6% 的准确率。Elwell 等人[17]主要进行自动语篇连接词论元识别,基于 Wellner 等人[16]的研究,添加了一些形态学、句法、词汇、语篇模式等特征,在连接词识别以及论元识别方面都获得了较好的结果。Wellner[18]用序列模型和排序方法进行语篇分析,给出了完整地进行语篇分析的步骤,在论元识别模块,识别论元的句法中心而不是整个论元,这样就使得论元的识别相对简单。该方法识别显式和隐式连接关系的准确率是 74.8%,正确句法树上 arg1 识别准确率是 80%,arg2 识别准确率是 91%,自动句法树上 arg1 识别准确率是 62%,arg2 识别准确率是 86%。上述文献在识别论元时都采用论元中心代替整个论元,使识别相对简单。Prasad 等人[19]利用范围进行浅层语篇结构分析,给出了一种方法自动识别连接词的论元。对于连接词的两个论元 arg1 和 arg2,因为 arg2 一般在连接词后面和连接词相邻,所以较好识别;arg1 比较难识别,故主要是识别 arg1。该方法采用基于句子的论元表示方法,将论元分为句子内部和句子之间两种情况进行处理。该方法主要是处理句子之间的情况,即 arg1 和 arg2 在不同的句子中,识别的结果是包含 arg1 的句子。文章采用了一系列与范围有关的过滤器过滤掉连接词前边的句子,降低选择空间,剩下的句子采用启发式的基于指代关系进行处理,从中选择合适的句子。结果表明识别 arg1 的效果有较大的提升,但是在识别论元时识别的是包含论元的句子,并不是真正的论元。

由以上分析可知,论元的识别中,识别 arg2 较容易,识别

arg1 较困难。目前论元的识别都是采用替代的方法,或者是将识别整个论元简化为识别论元中心,或者是识别包含论元的整个句子,这两种做法都不准确,应该探索有效的方法准确地识别论元,这样更有助于语篇关系的识别。

3.2.2 语篇关系识别

PDTB 中的语篇关系主要有显式和隐式语篇关系,显式语篇关系由于有连接词故较易识别,识别的准确率也较高。Pitler 等人[20]给出了一种识别显式语篇关系的分类器,显式关系由于有连接词故较易识别,总的来说连接词是没有歧义的,但是时间语篇关系有歧义。实验结果对显式语篇关系识别的准确率是 93.09%,总的识别准确率是 74.74%。另外,Pitler 等人分析得出有些关系的出现并不是随机的而是有规律的,这个发现表明全局序列语篇关系识别可以获得较好的结果,特别是对识别隐式关系帮助较大。Pitler 等人[21]使用句法特征消除显式语篇关系歧义,语篇连接词或短语如 once、since、on the contrary 等明显表示语篇之间的存在关系。但语篇处理中通常有两种歧义需要解决:a)一个词可以是语篇连接词也可以不是语篇连接词,如 once 可以是一个时间连接词,也可以是一个简单的词表示以前;b)一些连接词所表示的语篇关系也有歧义,如 since 可以表示时间关系,也可以表示原因关系。文献给出了一些句法特征(self category, parent category, left sibling category, right sibling category, right sibling contains a VP and right sibling contains a trace),这些特征对这两种歧义的解决都有所帮助。结果表明用最好的特征组合识别连接词是否表示连接关系的 F 值为 94.19%,说明显式关系的识别准确率较高。

隐式语篇关系由于没有指定的连接词,识别起来相对较难,相关研究主要有:Pitler 等人在文本隐式关系自动预测[22]中,给出了一系列实验自动识别隐式话语关系,即关系没有明显的语篇连接词,如“但是”“因为”。文献中使用了多个语言学的特征,包括极性信息、Levin 动词类别、动词短语长度、上下文、词汇方面的特点;所识别的隐式关系主要是四大类,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 contingency 的 F 值为 47.13%,expension 的 F 值为 76.42%,temporal 的 F 值为 6.16%,comparison 的 F 值为 21.96%。Lin 等人[23]识别 PDTB 中的隐式关系,采用词对、论元包含信息、依存信息、句法信息进行隐式关系识别,识别二级语篇关系,得到的准确率为 40.2%,结果比当时报告的最好结果高 14.1%。文献还分析了隐式关系处理所面临的四个挑战:歧义、推理、上下文和知识面。Wellner[18]在语篇关系识别模块对显式和隐式关系都进行识别,采用连接词、论元中心、上下文信息、时态、词性信息等多种特征进行识别,识别分为粗粒度、半粗粒度和细粒度三种情况。粗粒度识别分为四大类和其他(表示实体关系和无关系)共五类,识别的准确率为 68.88%;半粗粒度共识别九类关系,识别的准确率为 57.29%;细粒度识别的准确率为 51.36%。Louis 等人[24]使用实体特征进行隐式语篇关系分类,主要处理相邻句子之间的隐式语篇关系,采用与实体相关的特征如指代关系、语法角色、句法信息等分类隐式语篇关系,结果表明比原型系统采用的有歧义的特征要好。但是与实体相关的特征与传统的词汇特征相比,在隐式语篇关系识别上优势不太明显。

在语篇关系的识别上,显式关系的识别较容易,只需要解决连接词的识别和歧义问题。隐式关系的识别较困难,因为不存在明显的指示词,识别的准确率不高,有进一步提升的空间。

3.3 其他相关研究

3.3.1 结合 RSTDT 和 PDTB

Hernault 等人[25]提出利用特征向量扩充提高少量语篇关系分类的一种半监督方法。许多语篇结构分析采用全监督的机器学习方法,这种方法要求人为地提前标注一个训练语料,这是一个费时且代价高的工作,而没有标注的数据就比较多并且较易获得。Hernault 等人给出了一种关于语篇结构分析的半监督方法,采用以前使用过且报告有用的特征进行语篇关系学习,探索在无标注的文档上进行语篇关系识别的可能性,特别是改进类别例子较少的语篇关系性能。该方法采用同现矩阵进行特征扩充,在 RSTDT 和 PDTB 上都进行了实验,结果表明准确率和 F 值都有较大的提高。该文章首次提出关于出现频率较少的语篇关系的处理方法,这种方法对于语料较少的领域也比较有用。

Hernault 等人[26]采用结构学习的半监督语篇关系分类方法进行语篇关系识别。通常语篇关系语料中语篇关系标注的是一个通用的集合,但是对于特定应用,要求特定的语篇关系。标注一个新的语料库费时费钱,故提出了利用结构学习,采用半监督的方法进行语篇关系识别。这种方法对于语料标注较少的领域比较实用。

上述研究主要是采用半监督的方法进行语篇关系的识别,解决了某些领域语篇标注较少的情况,但方法的性能有待进一步的提升。

3.3.2 语篇结构分析应用

单纯的语篇结构分析意义不大,语篇结构分析的最终目的是为了应用。语篇结构分析的应用范围较广,举例如下:

1)自动摘要 Marcu[27]根据修辞结构理论和句子的主要部分信息来决定文本中最重要的信息,从《美国科学》中选择了五篇文档进行实验,结果表明使用语篇结构进行摘要抽取的准确率可达 70% 左右。Louis 等人[28]研究语篇指示在摘要选择中的作用,其目标是寻找语篇的哪个特殊方面在标注文本的重要性上起作用。在单文档摘要内容选择中,研究总结了语篇结构信息、图结构信息和语义语篇关系信息对摘要的作用。结果表明结构信息最有用,语义类别只在内容选择时提供限制但并不代表重要内容,但语义类别是结构信息的补充,可以提高系统性能。结构信息、语义信息、无语篇特征(包括句子长度、是否是段中首句或文中首句、位置信息等)的相互补充,使得判断文中重要句子的 F 值可达 44.3%,获得的摘要结果 rouge-1 可达 0.479。

2)指代消解 文献[2]给出几种连贯性关系的形式化定义,分析语篇关系同时可以得出指代关系。Webber 等人[29]讨论了许多副词短语在作语篇连接词的同时,也体现出存在语篇关系的相邻语篇单元之间也存在指代关系。

3)问答系统 Verberne 等人[30]给出一个基于语篇结构的 Why 问题回答方案,采用 RST 结构寻找 Why 问题的答案,利用语篇关系在 RSTDT 中生成 Why 问题的答案,实验召回率为 53.3%,MRR(mean reciprocal rank)为 0.662。如果答案不考虑世界知识,最大的召回率为 73.9%。结果表明语篇结果对于问答系统起重要作用,但提高召回率需要语言学的处理。Prasad 等人[31]给出了一个基于语篇方法的 Why 问题生成方法,分析 PDTB 中标注的因果关系在 Why 问题生成中的作用,实验表明 Why 问题语料中 71% 的内容和 PDTB 中标注的因果关系相关,说明 PDTB 标注的语篇关系对问答系统非常有用。

4)连贯性　Lin 等人[32]用语篇关系自动评估文本连贯性,给出了一个表示和评估文本连贯性的模型。实验结果表明利用语篇关系可以提高 Barzilay 的连贯性模型,减少错误率。

3.3.3　中文语篇结构分析研究情况

中文语篇结构分析的研究主要是扩展谓词—论元的概念。首先建立中文语篇树库,在中文语篇树库中连接词被认为是带有论元的谓词,语篇连接关系可以是从属关系、并列关系或照应状语表达。目前正在开发的是 Chinese Penn discourse treebank。

4　研究趋势

目前语篇结构分析的研究越来越与实际应用相结合,用来解决实际的问题,如自动文摘、极性判断、连贯性、指代消解等,但目前所使用的语篇关系都是浅层的、较好识别的关系,深层语篇结构分析的准确性还有待提高。只有提高语篇分析的准确性才能更好地理解语篇所要表达的内容,和其他自然语言处理技术相结合,提高自然语言处理的总体水平。

由第 3 章所述可知,语篇结构分析是一个复杂的工作,要求至少完成这样几步:识别句子是否存在语篇关系,识别存在语篇关系的基本语篇单位,分类得出具体的语篇关系类别。

识别句子是否存在语篇关系,有连接词的情况较好识别,没有连接词的情况较难处理,但不能识别出是否存在语篇关系则下面的工作就很难进行,所以这个问题必须解决。目前这方面的研究力度较小,需要进一步的研究。

目前 RSTDT 中识别基本语篇单位准确率较高,PDTB 中因为语篇单位跨度有大有小,主要采用替代的方法,或识别基本语篇单位中心,或识别包含基本语篇单元的句子,都是尽量地想让问题简化。问题简化识别起来方便,但与实际语料中标注的情况会有出入,所以最终还要识别整个论元。

在语篇关系识别研究上,显式的语篇关系识别准确率较高,只需解决连接词歧义的问题,隐式语篇关系识别率较低,还要进一步研究提高。只有较准确地识别语篇关系,才能在实际应用中发挥作用。

自 PDTB 2.0 于 2008 年发布以来,由于其包含文档较多、覆盖面较广、标注的语篇关系类别层次较多,适用于不同的应用,目前成为语篇关系识别的主要语料。

在 PDTB 的研究中,目前识别句子是否存在语篇关系的研究较少,一般是采用语料中标注好的关系进行研究,自动判断句子之间是否有关系是语篇关系研究的内容之一,所以需要研究。目前,arg2 的识别准确率较高,arg1 的识别准确率较低。arg2 和 arg1 的表示方法主要采用论元中心或论元包含的表达方式,需要探索新的论元边界识别方法,因为在实际应用中用到的主要是完整的论元信息。在语篇关系识别研究中,显式语篇关系识别较容易,准确率较高;隐式语篇关系由于没有连接词,识别较困难。在语篇关系具体类别上,第一层的四种关系识别较容易,第二层和第三层由于粒度较细,识别起来困难较大,这也是研究的热点和难点之一。

汉语语篇结构分析起步较晚,目前的研究较少,随着 CDTB 语料库的完善,汉语语篇结构分析的研究将迎来一个新的阶段。在汉语语篇结构分析中,可以参考英语语篇结构分析的结果,采用类似的方法进行,但由于汉语和英语的不同,需要探索针对汉语语篇结构分析的特殊方法。

5　结束语

本文系统全面地介绍了英语语篇结构分析的发展历史和研究现状,在分析当前制约英语语篇结构分析发展原因的基础上给出了目前该领域的研究趋势,为英语和汉语语篇结构分析研究工作的进一步展开做了基础性的工作。

参考文献:

[1] HALLIDAY M, HASAN R. Cohesion in English[M]. [S. l.]: Longman, 1976.

[2] HOBBS J R. Coherence and coreference[J]. Cognitive Science, 1979, 3(1): 67-82.

[3] HOBBS J R. Information, Intention, and structure in discourse: a first draft[C]//Proc of NATO Advanced Research Workshop on Burning Issues in Discourse. 1993: 41-66.

[4] MANN W C, THOMPSON S A. Rhetorical propositions in discourse[J]. Discourse Processing, 1986, 9(1): 57-90.

[5] MANN W C, THOMPSON S A. Rhetorical structure theory: toward a functional theory of text organization[J]. Text, 1988, 8(3): 243-281.

[6] WOLF F, GIBSON E. Representing discourse coherence: a corpus-based analysis[J]. Computational Linguistics, 2005, 31(2): 249-287.

[7] CARLSON L, MARCU D, OKUROWSKI M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory[M]// Current Directions in Discourse and Dialogue. San Francisco: Kluwer Academic Publishers, 2003: 85-112.

[8] PDTB Research Group. The Penn discourse treebank 2.0 annotation manual, IRCS-08-01[R]. Philadelphia: University of Pennsylvania, 2008.

[9] SORICUT R, MARCU D. Sentence level discourse parsing using syntactic and lexical information[C]//Proc of Human Language Technology and North American Association for Computational Linguistics Conference. 2003: 149-156.

[10] Le HUONG T, GEETHA A, CHRISTIAN H. Automated discourse segmentation by syntactic information and cue phrases[C]//Proc of International Conference on Artificial Intelligence and Applications. 2004.

[11] SUBBA R, EUGENIO B D. Automatic discourse segmentation using neural networks[C]//Proc of the 11th Workshop on the Semantics and Pragmatics of Dialogue. 2007: 189-190.

[12] TOFILOSKI M, BROOKE J, TABOADA M. A syntactic and lexical-based discourse segmenter[C]//Proc of ACL-IJCNLP Conference Short Papers. 2009: 77-80.

[13] HERNAULT H, BOLLEGALA D, ISHIZUKA M. A sequential model for discourse segmentation[C]//Proc of the 11th International Conference on Computational Linguistics and Intelligent Text Processing. 2010: 315-326.

[14] Le HUONG T, GEETHA A, CHRISTIAN H. Generating discourse structures for written texts[C]//Proc of the 20th International Conference on Computational Linguistics. 2004: 329-335.

[15] DuVERLE D A, PRENDINGER H. A novel discourse parser based on support vector machine classification[C]//Proc of the 4th Annual Meeting of ACL and the 4th Conference on Natural Language Processing. 2009: 665-673.

[16] WELLNER B, PUSTEJOVSKY J. Automatically identifying the arguments of discourse connectives[C]//Proc of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007: 92-101.

(下转第 2027 页)

(上接第 2023 页)

[17] ELWELLR, BALDRIDGE J. Discourse connective argument identification with connective specific rankers[C]//Proc of IEEE International Conference on Semantic Computing. 2008:198-205.

[18] WELLNER B. Sequence models and ranking methods for discourse parsing[D]. Waltham: Brandeis University,2009.

[19] PRASAD R, JOSHI A K, WEBBER B L. Exploiting scope for shallow discourse parsing[C]//Proc of the 7th International Conference on Language Resources and Their Evaluation. 2010:2076-2083.

[20] PITLER E, RAGHUPATHY M, MEHTA H, *et al*. Easily identifiable discourse relations, MS-CZS-08-24[R]. Philadephia: University of Pennsylvania,2008:683-691.

[21] PITLER E, NENKOVA A. Using syntax to disambiguate explicit discourse connectives in text[C]//Proc of ACL-IJCNLP Conference Short Papers. 2009:13-16.

[22] PITLER E, LOUIS A, NENKOVA A. Automatic sense prediction for implicit discourse relations in text[C]//Proc of the 4th Annual Meeting of ACL and the 4th Conference on Natural Language Processing. 2009:683-691.

[23] LIN Zi-heng, KAN Min-yen, NG H T. Recognizing implicit discourse relations in the Penn discourse treebank[C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2009:343-351.

[24] LOUIS A, JOSHI A, PRASAD R,*et al*. Using entity features to classify implicit discourse relations[C]//Proc of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2010:59-62.

[25] HERNAULT H, BOLLEGALA D, ISHIZUKA M. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension[C]//Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2010:399-409.

[26] HUGO H, DANUSHKA B, MITSURU I. Semi-supervised Discourse Relation Classification with Structural Learning[C]//CICLing 2011, Part I, LNCS 6608,2011:340-352.

[27] MARCU D. From discourse structures to text summaries[C]//Proc of the ACL/EACL Workshop on Intelligent Scalable Text Summarization. 1997:82-88.

[28] LOUIS A, JOSHI A, NENKOVA A. Discourse indicators for content selection in summarization[C]//Proc of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2010: 147-156.

[29] WEBBER B, STONE M, JOSHI A, *et al*. Anaphora and discourse structure[J]. Computational Linguistics,2003,29(4):545-587.

[30] VERBERNE S, BOVES L, COPPEN P A, *et al*. Discourse-based answering of why-questions[J]. Traitement Automatique des Langues,2007,47(2):21-41.

[31] PRASAD R, JOSHI A. A discourse-based approach to generating why-questions from texts[C]//Proc of Workshop on the Question Generation Shared Task and Evaluation Challenge. 2008.

[32] LIN Zi-heng, NG H T, KAN Min-yen. Automatically evaluating text coherence using discourse relations[C]//Proc of the 49th Annual Meeting of the Association for Computational Linguistics. 2011:997-1006.