doi:10.6043/j.issn.0438-0479.201805011

·综 述·

面向中文抽象语义表示的复句研究综述

魏庭新[1,2],曲维光[2,3,4*],宋 丽[2],戴茹冰[2]

(1.南京师范大学国际文化教育学院,2.南京师范大学文学院,3.南京师范大学计算机科学与技术学院,江苏 南京 210097;4.福建省信息处理与智能控制重点实验室(闽江学院),福建 福州 350121)

摘要:抽象语义表示(AMR)是一种新型的句子语义表示方式.中文 AMR 在英文 AMR 的基础上,针对汉语特点,增加了复句逻辑语义关系的表示.中文 AMR 以句子为基本标注单位,以层次结构树形式表示各分句间的逻辑关系.由于允许论元共享,因此在树结构基础上形成图结构,从而对复句的语义表示更加完整全面.为了进一步研究中文 AMR,对目前复句关系研究现状、复句及篇章关系资源的建设进行了综述,指出目前研究存在的问题,并提出将来工作研究的方向.

关键词:中文抽象语义表示;复句;篇章关系

中图分类号:TP 391.1 **文献标志码**:A **文章编号**:0438-0479(2018)06-0849-10

语义是语言符号的意义,是说话人通过语言形式最终想传达的信息.自然语言处理的一个重要研究目标就是通过外在的词汇、句法等语言形式来实现对语义的理解,因此对语义的解析一直是自然语言处理的热点之一.近年来在词汇语义标注、语义角色标注、共指消解等方面都取得了长足的进展,然而在整句逻辑语义表示和解析方面仍不甚理想.2013 年美国宾夕法尼亚大学的语言数据联盟(LDC)连同南加州大学、科罗拉多大学等共同提出了一种新型的语义表示语言,即抽象语义表示(abstract meaning representation AMR)[1],它采用图结构来表示一个句子的语义.这种表示方法在保留了句子树形主干结构的同时,使用有向无环的逻辑图结构,实现了对句子中论元共享现象的表示.同时它还允许添加原句缺省的概念节点,以更好地表示其隐含意义[2].因此一经推出,就受到国内外学者的高度关注,引发了一股研究 AMR 的热潮.目前 AMR 的标注和解析都是以句子为基本单位进行的,而自然语言中的句子根据复杂度可以分为单句和复句,两者在句法、语义上有着巨大的不同.随着 AMR 研究的深入,迫切需要对句子本身的逻辑语义进行更深入细致的研究和挖掘.因此本文中对中文 AMR(CAMR)的复句研究进行综述和分析,对 CAMR 在复句处理方面的工作提出了研究展望.

1 理论研究背景

1.1 单句、复句与篇章

篇章是一定语境中表示完整语义的一系列句子或语段构成的语言单位,而句子是篇章的基本单位.根据内部结构不同,句子可以分为单句和复句.如果一个句子是由两个及以上的意义相关的句子组成的,彼此分立,互不作为句子成分,这样的句子称为“复句”[3].20 世纪 50 年代语言学界曾经有一场关于单复句的广泛讨论,虽然各家争鸣,意见不同,但最后也达成了一些共识:单句和复句结构上有着本质的不同,其中一个主要不同之处在于,复句除了分句本身的语义,还包含分句之间的逻辑语义.胡金柱等[4]形象地将之表示为:复句语义=逻辑语义+$\sum_i$分句 i 语义.同时他还指出,复句是连接分句与篇章的桥梁.Mann 等[5]的修辞结构理论(rhetorical structure theory,RST)认为,篇章结构具有组织性、连贯性、层级性、层级同质性等特

收稿日期:2018-05-09 **录用日期**:2018-10-04

基金项目:国家自然科学基金(61772278,61472191);福建省信息处理与智能控制重点实验室开放基金(MJUKF201705)

* **通信作者**:wgqu_nj@163.com

引文格式:魏庭新,曲维光,宋丽,等.面向中文抽象语义表示的复句研究综述[J].厦门大学学报(自然科学版),2018,57(6):849-858.

Citation:WEI T X,QU W G,SONG L,et al. A survey on the study of compound sentences with Chinese abstract meaning representation[J]. J Xiamen Univ Nat Sci,2018,57(6):849-858.(in Chinese)

征.对于复句与篇章的关系,徐赳赳[6]将复句研究与RST比较之后认为,复句已经进入篇章研究的范围,特别是多重复句,篇章的特征更明显.他发现汉语的复句理论与RST在研究的基本单位、研究对象上有很多相似之处.可以说,复句关系和篇章关系是同构的,篇章语义关系几乎都可以在复句语义关系中找到.搞清楚复句语义关系,篇章语义关系便能够迎刃而解.因此,无论是对句子进行句法分析,还是解析篇章语义,对复句进行研究和处理都是十分必要的.

1.2 关于复句语义关系的研究

关于汉语复句的分类,学界并无统一标准,各家均有自己的主张.比较有代表性的有以下几种:黄伯荣等[7]采取两分法(以下简称黄廖二分法),将复句分为联合复句和偏正复句两大类,联合复句下辖并列、顺承、解说、选择、递进5个小类,偏正复句又分为转折、条件、假设、因果、目的复句.邢福义[8]采取三分法,把复句分为因果、并列、转折3大类,因果类下分因果、推断、假设、条件、目的小类,并列类下分并列、连贯、递进、选择等,转折类包括转折、让步和假转等.胡明扬等[9]则根据是否有形式标志,将复句分为有关联词复句和无关联词复句,无关联词复句又分为意合句、流水句和排比句等.在英语中,由于语言本身的特点,鲜有专门针对复句关系的研究,多数研究从篇章层面来考虑主从句、复句、句群之间的语义关系.如宾州树库体系[10]主要考虑句间语义关系,将篇章关系分成了因果、比较、扩展、时序4大类.而RST[5]则从一致性、连贯性、主次关系等角度考虑篇章各层次语块间的修辞关系,总结了包括证明、条件、解释、对立等关系在内的24种关系.

2 CAMR复句处理的方法及特点

2.1 AMR对句子的语义标注及特点

2016年LDC公布了英文《小王子》的AMR标注语料,2017年又发布了AMR2.0版本[11],内含来源于网络论坛、博客、华尔街日报、新华日报英文版等在内的39 260个句子的AMR标注.AMR在标注句子语义时有这样2个特点:

1)以句子为基本单位,对句子的整体语义做抽象表示.切分后的AMR句子基本为单句,或者是带有定语从句、主语从句或宾语从句的复合句.对于复句,AMR的处理方法是将之进一步切分为单句,不以复句为单位进行处理,也不处理切分后的句间关系.

2)只允许一个句子有一个根节点,对于带有从句等结构作修饰成分的复合句,则根据修饰成分与中心语的语义关系将其标记为一个论元附着在相应节点上.

2.2 汉语和英语的不同

从类型学来说,汉语和英语是两种非常不同的语言.汉语缺乏形态变化和形态标志,重意合;而英语形态结构完备,重形合.王力[12]指出,就句子的结构而论,西洋语言是法治的,中国语言是人治的.这些特点反映在句子层面,使得英语多长句,汉语多短句;英语多从句,汉语多分句;汉语还有独特的流水句.英语句子虽长,各种语义角色能以从句形式依附在主干结构上,这与AMR的分析方法是比较契合的.然而汉语由于缺乏形态标记,多用分句来表达复杂语义,多个分句共同完成一个完整语义的表达,且分句句法成分常常承前省略.如果还按照AMR处理复句的方法,将分句切分,势必会使得句子语义表示不完整,因此如果想在汉语上使用AMR,必须根据汉语的特点,对AMR的标注方法做出相应调整.

2.3 CAMR的复句处理方法

2016年,Li等[13]基于AMR框架结构,同时考虑了汉语与英语的差异,初步建立了一套中文抽象语义的表示方法和标注规范.标注规范针对中英文的差异做了很多调整,如对汉语特有的量词、把字句、被字句等汉语特殊句式等做了相应规定.在句子处理层面,对于复句,没有采用英文AMR直接切割为单句的做法,而是将构成复句的句间语义关系作为该句的根节点,语义关系所涉及的分句作为该语义关系的论元arg1,arg2.同时根据汉语特点,并借鉴中文语料树库(Chinese discourse treebank,CDTB)标注汉语篇章关系的方法[14-15],在标注时增加了10类复句关系,包括:并列、因果、条件、转折、时序、选择、让步、解释、目的、递进.例如,"孔子学生赎一奴,却不报账,人人夸学生高尚."的CAMR可表示为:

```
(n0 / causation
    :arg 1 (n1 / contrast
        :arg1 (n2 / 赎-01
            :arg0 (n3 / 学生
                :poss (n4 / person
                    :name (n5 / name :op1 孔子)))
            :arg1 (n6 / 奴
                :quant (n7 / 1)))
        :arg2 (n8 / 报账-01
            :polarity (n9 / -)
```

```
        :arg0 n3))
    :arg2 (n10 / 夸-01
        :arg0 (n11 / 人)
            :mod (n12 / every)
        :arg2 (n13 / 高尚-01
            :arg0 n3)
        :arg1 n3))
```

可以看到,CAMR 将句间语义关系“causation(因果)”作为复句根节点,该关系所涉及的两个句子作为其论元,然后再分别对两个论元进行表示,对于仍然包含一个复句语义关系的论元 arg1,则继续将语义关系“contrast(转折)”作为根节点,所涉及两个分句作为其论元.另外 CAMR 标注了分句之间的层次结构,以缩进的形式清楚地呈现出来.

2.4　CAMR 复句语义标注方法与其他体系比较

2.4.1　CAMR 复句语义关系与其他体系比较

本文中将 CAMR 的复句语义关系与目前语言学界广泛使用的黄廖二分法[7]、邢福义的三分法[8]以及清华汉语树库[16]的句间语义关系、苏州大学汉语篇章结构语料库[17]的篇章关系分类方法进行了比较,结果如表 1 所示.

表 1　CAMR 与不同体系复句语义分类比较

Tab. 1　Compound sentences′ semantic classifications of CAMR and other ontologies

CAMR	黄廖二分法		邢福义三分法			清华汉语树库	汉语篇章结构语料库			
	联合	偏正	因果	并列	转折		因果	并列	转折	解说
并列	并列	转折	因果	并列	转折	并列	因果	并列	转折	解说
因果	选择	条件	条件	递进	让步	因果	推断	顺承	让步	总分
条件	递进	因果	目的	选择	假转	条件	假设	递进		例证
转折	顺承	目的	推断	连贯		转折	目的	选择		评价
递进	解说	假设	假设			递进	条件	对比		
选择						选择	背景			
让步						目的				
解释						连贯				
目的						假设				
时序						解注				
						流水				

可以看到,尽管各种分类方法对语义关系的分层不同,但均包含并列、因果、条件、转折、递进、选择等几种具体语义关系,CAMR 吸收了这些学术界普遍认同的关系.由于假设关系从逻辑上来说也是一种条件关系,因此 CAMR 将之归并入条件关系.解释关系是否是一种主要的复句关系语言学界意见并不一致,时序关系是传统汉语复句关系不太注重分析的,但宾州篇章树库、修辞结构理论篇章树库(rhetorical structure theory discourse treebank,RST-DT)[18]等篇章关系语料库普遍采用这两种语义关系,说明其对于揭示复句中分句间的逻辑语义有着重要的作用,因此 CAMR 也吸收了这两种语义关系.特别是对于汉语特有的流水句,时序关系可以比较精准地解释各分句间的语义关系.如“开放以后,大陆富裕了,香港人发现,赚钱不是自己的独门绝活.”的 CAMR 可表示为:

```
(n0 / temporal
    :arg1 (n1 / 开放)
    :arg2 (n2 / 富裕-01
        :aspect (n3 / 了)
        :arg0 (n4 / 大陆))
    :arg3 (n5 / 发现-01
        :arg0 (n6 / 香港人)
        :arg1 (n7 / 绝活
            :mod (n8 / 独门)
            :poss n6
            :domain (n9 / 赚-01
                :arg1 (n10 / 钱))
            :polarity (n11 / -))))
```

可以看到,相比其他几种语义关系,表示“大陆富裕了”和“香港人发现”两个分句间语义最确切的就是时序关系.

2.4.2　CAMR 复句标注单位与其他体系的比较

CAMR 复句标注的对象是经过 Stanford CoreNLP 切分后结构为复句的句子,因此 CAMR 复句标注的

基本单位是具有独立表述功能的最小单句,不仅包括了由逗号标记的分句,还包括紧缩复句中有独立表述功能的短语段,如果含有大于分句的语言片段则继续切分. PDTB(Penn discourse treebank)在标注时面向篇章关系,所以标注单位是句子甚至是句群,与 CAMR 相比颗粒度较粗. RST-DT 在短语级、句子级、篇章级都进行切分和标注,短语级的标注更多揭示的是句内谓词论元关系,而非篇章关系. 中文篇章关系分析如哈尔滨工业大学篇章关系语料库(HIT-CDTB)也是以句群为切分单位,并没有细化到最小分句. 几种体系的标注单位比较如表 2 所示.

表 2 CAMR 与 PDTB,RST-DT,HIT-CDTB 标注单位对比表

Tab. 2 Unit comparison of CAMR, PDTB, RST-DT and HIT-CDTB

标注单位	CAMR	PDTB	RST-DT	HIT-CDTB
整体单位	句子	篇章	篇章	篇章
基本单位	小句	从句/句子/句群	基本篇章单元(短语/从句/句子)	句子/句群
例句	arg1:我亲属刚刚打了一场官司,arg2:[arg1:按照法律是完全没有争议和责任的,arg2:地方法院却判了我亲属赔人家十几万为之].	arg1:The average 6-month bill was sold with a yield of 8.04%,up from 7.90%. arg2:The average 3-month issue rose to 8.05% from 7.77%.	Her sister, Cynthia, wishes (EDU) Toni had a different job(EDU).	arg1:超过两百名乘客被卡在翻覆的车厢内动弹不得,哀嚎声四起. arg2:救援人员动用了 6 台起重机吊正翻覆的车厢,并使用大型破坏机具和吹管切割车厢,才陆续救出被困在车厢里的乘客.

注:EDU(elementary discourse unit)即基本篇章单元.

可以看到,PDTB 和 HIT-CDTB 的标注体系中,篇章关系的论元仍可以包含多个小句,RST-DT 的 EDU 可以是单句的一个部分;而 CAMR 的最小单位则是句子的最小分句.

2.4.3 CAMR 复句层次标注与其他体系比较

目前宾州篇章树库、清华汉语树库、汉语复句语料库等都只关注语义关系,不对层次进行标注. 进行层次标注的有 RST-DT 和汉语篇章结构语料库,这些研究均采用树结构来对篇章单元之间的层次关系进行描述,而 CAMR 是将句中所有概念的语义抽象出来,对复句层次划分采用树结构的同时,允许论元共享,因此形成图结构. 如"问题不是出在中国而是出在美国."的 CAMR 可表示如下:

```
(n0 / contrast
    :arg1 (n1 / 出-07
        :polarity (n2 / -)
        :arg1 (n3 / 问题)
        :arg0 (n4 / country
            :name (n5 / name :op1 中国)))
    :arg2 (n6 / 出-07
        :arg1 n3
        :arg0 (n7 / country
            :name (n8 / name :op1 美国))))
```

可以看到,CAMR 将该复句分为转折关系的两个分句之外,还指出后一分句的根节点与前一分句的根节点共享 arg1"问题". 这样,CAMR 的复句语义不仅含有句间逻辑语义关系,还将各分句缺省的论元补充完整,相较其他篇章关系分析体系只关注句间语义,这也是 CAMR 在句子语义表示方面的一大优势.

2.5 CAMR 复句研究任务

要做好 CAMR 中复句的自动标注和解析工作,就要做好以下子任务:1) 复句语料库的建设. 2) 复句语义关系识别. 根据是否含有关系词可将复句分为有标记复句和无标记复句. 对于有标记复句,要做的是关系词的识别,然后根据关系词语义来进行复句关系的识别;对于无标记复句,需要做的则是隐式语义关系的判定. 3) 复句结构层次的生成. 与一般篇章处理任务不同,CAMR 在判断语义后还要对各部分的语义进行层次判定和生成,从而生成最终的逻辑语义关系树.

3 研究现状

目前专门针对复句语义关系判定、解析的研究很少,大部分研究都是在篇章层面开展的,不过仍可以作为对复句研究的借鉴. 下面从资源建设、分句切分、关系识别、结构树生成等几方面对目前的研究现状进

行介绍.

3.1　资源建设

目前关于复句语义关系的语料库资源非常少,除了在建的 CAMR 语料库(http://www.cs.brandeis.edu/~clp/camr/camr.html)之外,只有华中师范大学汉语复句语料库[19],另外还有一些篇章关系语料库,如 PDTB(http://www.seas.upenn.edu/~pdtb/)、RST-DT(http://www.isi.edu/~marcu/discourse)、HIT-CDTB(http://ir.hit.edu.cn/hit-cdtb/index.html)、苏州大学汉语篇章结构语料库和清华汉语树库等可供借鉴.

CAMR 语料库:由美国布兰迪斯大学和南京师范大学共同开发,目前已标注 1 562 句中文《小王子》[20]及 10 325 句中文树库(Chinese treebank,CTB)中的网络语料,其中含有复句关系的有 7 899 句.

汉语复句语料库:华中师范大学开发的汉语专用语料库,语料主要来自《人民日报》和《长江日报》,同时还有一部分现当代文学作品,共收有标复句 80 万句.标注内容包括关系词类别、关系词连接项功能、分句层次、复句句式类别等.在复句语义关系体系上,采用了邢福义的三分法,将复句语义关系分为因果、转折、并列 3 大类,每大类下又各分小类,共 12 小类.该语料库目前尚未对外公布.

宾州篇章树库:是 LDC 2006 年发布的标注篇章关系的语料资源,2008 年发布了 2.0 版,内容来源为华尔街日报(WSJ)的 2 300 多篇文章,是目前最大、使用最多的篇章关系语料库.它参照 Propbank 的标注方法,将篇章中的文本片段标记为(连接词、论元)结构,其中连接词是联系上下文本片段的关系词,被连接的两个文本片段被标记为 Arg1、Arg2,它将篇章间的语义关系分为 4 种:显式/隐式连接关系、基于实体的关系(EntRel)、词汇替代关系(AltLex)、没有关系(NoRel).其中显式/隐式关系的语义体系又根据粒度不同分为 3 层,第 1 层 4 类,第 2 层 16 类,第 3 层 23 类.

RST-DT:是由 LDC2002 年发布的针对篇章修辞结构标注的语料资源,该语料库基于 Mann 等 1988 年提出的修辞结构理论建立,将篇章文本进行切分,形成独立且能表达一定语义的 EDU,并为篇章间语义定义了多种关系.它与 PDTB 的不同之处在于,它区别了篇章单元前后的主次关系,并将这种结构关系分为单核和多核;将同一篇章内篇章单元间的修辞关系层次划分出来,层层叠加,最终形成修辞结构树.该语料库规模较小,只标注了华尔街日报 385 篇英文文章.

HIT-CTDB:由哈尔滨工业大学社会计算与信息检索研究中心开发,语料来自 OntoNotes4.0 上的 525 篇中文文本.针对每篇文本,均标注了 3 种关系:分句篇章关系、复句篇章关系和句群篇章关系.标注采用 PDTB 标注体系,但篇章关系根据汉语特点做了相应调整,共分为时序、因果、条件、比较、扩展和并列 6 种关系[21].

汉语篇章结构语料库:由苏州大学开发,采用树的形式来表示汉语的篇章结构,每个段落构建一棵篇章结构树,标注了篇章中语义关系、连接词、中心、层次等信息.在语义关系分类上,该语料库将篇章关系分成 4 个大类、17 个小类,其中 4 个大类分别为因果、并列、转折和解说关系.

清华汉语树库(Tsinghua Chinese treebank):由清华大学开发,语料主要来自汉语平衡语料库,有文学、新闻、学术、应用等 4 种文体.这个语料库不仅标注复句语义关系,还标注了词类、短语结构、功能等多个信息.该语料库没有标注关系词,将复句语义关系直接分为并列、连贯、递进、选择、因果、目的、假设、条件、转折、注解、流水 11 种关系.

3.2　分句的自动切分

复句常常由多个分句组成,要解析分句间的语义关系,首先要解决分句的边界识别问题,即语义关系涉及的论元边界问题.在基于 RST-DT 的研究中,关于 EDU 识别的研究较多,如 Soricut 等[22]采用概率模型,利用句子的句法和词汇特征进行句子级别的 EDU 识别和篇章结构树的构建,在自动句法树上取得了 F 值为 83.1% 的识别效果;Hernault 等[23]等使用支持向量机(SVM)模型,利用句子的句法、结构、词汇等特征,在 EDU 识别任务上 F 值达到了 93.8%,效果良好.Lin 等[24]在 PDTB 上做的论元识别 F 值达到 82.6%.然而在汉语中,由于逗号除了做分句间隔之外还常常用作语气停顿标志,导致很多非分句片段的产生.要判定复句语义关系,首先要排除这些非分句短语片段的干扰.一些学者对此展开了研究,如洪鹿平[25]使用 SVM 分类器判断逗号前的文本片段是否为分句;胡金柱等[26]利用规则和聚类分析的方法对复句中的短语字段进行自动识别,准确率达到 92.1%.这些研究仅进行是否为分句的判断,对于非分句的归属等后续问题则没有深入研究.

3.3　复句关系的自动识别

3.3.1　显式关系识别

1) 关系词识别

复句中的显式关系指的是包含关系词的复句所

表示的逻辑语义关系.英语中的关系词大部分是非歧义的[27],因此只要识别出关系词,基本就可以推断出其表示的语义关系.对于有歧义的关系词,Pitler 等[28]使用词汇和句法特征来判断其是否为篇章关系词,准确率可以达到 96.26%,F 值达 94.19%,Lin 等[24]在此基础上抽取了词性、上下文等特征来构建其关系词分类器,最终准确率达到 97.25%,F 值达到 95.36%.与英语相比,汉语篇章中关系词的语法性质和词性分布更加复杂.李艳翠等[29]指出,汉语中的关系词不限于传统连词,还有介词、副词等诸多语法类型.胡金柱等[30]建立了一个复句关系词库,将复句中的关系词分为 3 类,第 1 类为语义单一型典型关系词,如"因为、所以"等,这些词能够固定地表示分句间的某种语义关系;第 2 类为语义多样型非典型关系词,如"就、才、也"等副词,可以兼表几种语义关系;第 3 类为语义单一型非典型关系词,如"别管、怪不得、谁知道"等形式上处于实义短语与关系词的共存状态.因此,汉语中关系词消歧任务比英语更加复杂和艰巨.李艳翠等[31]利用词的词汇、句法、位置特征使用决策树分类器在清华树库上进行是否为关系词的识别,在不带功能标记的词上达到了 92.1%的准确率,但该研究只识别单个关系词,而汉语中关系词常常是成对成组出现的.针对这一问题,杨进才等[32]使用贝叶斯模型对关系词的特征集合进行训练和测试,将基于统计过程的结果转换为规则,在汉语复句语料库上取得了 95.4%的准确率.该研究实验数据较小,只验证了 15 组关系词在 1 000 句上的准确率.总的来说,目前汉语关系词识别效果较好,但研究多是着眼于典型关系词,对于非典型关系词的识别较少.

2) 显式语义关系判定

在连接关系识别领域,Pitler 等[28]仅使用关系词特征,在 PDTB 分类体系下将篇章语义分成因果、比较、时序和扩展,取得了 93.9%的准确率.Lin 等[24]在特征中加入了关系词,上下文等特征,在自动句法树上取得了 86%的准确率.汉语中由于关系词歧义情况较为复杂,目前取得的效果较英文稍差.李艳翠等[31]在 PDTB 分类体系下使用最大熵分类器对连接词语义进行分类,4 分类的准确率仅有 78.9%,F 值也仅有 69.3%.张牧宇等[33]使用极大似然估计法,利用关系词特征进行关系分类,在因果、条件、比较关系上都取得比较好的效果,准确率均超过 95%,但在并列关系上效果较差,准确率只有 63.6%.以上研究都是在 4 大类分类上实验,没有将语义关系进一步细分为小类.杨进才等[34]对于只有部分分句含有关系词的非充盈态有标复句计算分句核心词的语义相关度,作为判断复句语义关系的依据,准确率达到了 89%,但没给出各类别的准确率.可以看到,汉语显式语义关系识别仍有一定的提高空间.

3.3.2 隐式关系识别

显式复句关系词可以作为判定语义关系的强力标志,而不含关系词的隐式关系判定则给复句语义关系识别带来巨大挑战,也是目前篇章关系研究领域的热点.

1) 基于特征的方法

Marcu 等[35]抽取论元的词对信息,利用互联网抽取大量词对信息实例,并将其中的关系词移除构建一个隐式关系语料库,然后使用贝叶斯分类器对隐性语义关系进行识别.Pitler 等[36]则将词的情感特征、动词类别、动词短语长度、情态、上下文和词汇特征等用于篇章关系识别,在 PDTB 4 类语义关系分类任务上,各类特征的使用对于结果的 F 值提升都有明显作用.Lin 等[37]使用前后论元信息、词对信息、论元内部成分和依存句法信息作为特征,利用最大熵分类器,在 PDTB 第 2 层 11 类语义关系上进行识别,取得了 40%的准确率,比 baseline 提高了 14.1%.Louis 等[38]尝试将文本中的指代信息以及指代词的句法结构和特征用于隐性语义关系的识别,效果虽较 baseline 有提升,但比传统利用词法特征的方法仍然相差较多.Rutherford 等[39-40]针对有些显性关系移除关系词后意义改变不能用于构造隐性关系的问题,通过计算关系词的省略率来选出合格的关系词论元对,进而扩大训练数据集,提升了识别效果,在 PDTB 4 分类上准确率达到 40.5%.车婷婷等[41]挖掘词级和短语级的功能连接词,建立功能连接词的概念模型与篇章关系的映射体系,实现隐式篇章语义关系的推理,虽然结果取得了不错的效果,准确率达 53.84%,但是只比全部标为最大类别扩展关系的 baseline 准确率高 0.1%,这也说明目前隐式篇章关系识别的难度.

在汉语隐式篇章关系研究方面,张牧宇等[33]基于有指导方法的关系识别模型,利用核心动词、极性特征、依存句法特征、句首词汇特征等,对因果、比较、扩展、并列 4 类关系进行分类,结果只有扩展关系的识别效果不错,F 值达到 72.3%,其他 3 类效果不佳,比较关系的 F 值最低,只有 16.2%.孙静等[42]利用上下文特征、词汇特征、依存树特征,采用最大熵分类法对因果、并列、转折、解说 4 大类关系进行识别,总准确率为 62.15%,但除了并列类效果很好之外,其他 3 类效果都不佳,特别是转折类完全没有识别出来.李国

臣等[43]利用汉语框架语义网识别 11 种篇章语义关系,结果显示只有属于关系识别效果较好,准确率超过 70%,其他关系效果都不尽理想,均低于 40%.

可以看到,无论是在英语还是汉语中,传统基于特征的方法准确率都不高,扩展或并列类准确率较高的原因是自然语言中这类语义关系本身占比就较大,若剔除这个因素,准确率可能还要更低.想要提高性能,必须表征句子更深层的语义关系.

2) 基于神经网络的方法

随着近些年神经网络研究的兴起,学者们发现相比于传统方法使用浅层特征易于丢失文本序列、结构等重要信息,使用词嵌入(word embedding)对句子进行表示更能获取句子深层的语义信息.在机器翻译、阅读理解等领域取得卓越效果之后,一些学者也开始将神经网络用于隐式篇章关系的识别.Ji 等[44]最早将神经网络技术应用于篇章隐式关系,他们用循环神经网络(recurrent neural network,RNN)对句子的论元及实体进行编码,在 PDTB 4 类语义分类任务中将准确率提升到了 43.56%.Zhang 等[45]则是使用了只有一个隐藏层的浅层卷积神经网络(SCNN)在 PDTB 上进行隐式关系识别,并在 4 个关系分类任务中的 3 个(因果、扩展、时序)上取得了优于基于 SVM 方法的结果.Liu 等[46]使用双向长短期记忆网络(Bi-LSTM)将隐式关系中的论元编码,同时模仿人类重复阅读习惯,引入了多重注意力(multi-attention)机制,对隐式篇章关系进行识别,在 PDTB 4 类关系的分类中准确率和 F 值分别为 57.57%和 44.95%.Li 等[47]对论元、句子和段落都进行分布式语义表示并将之组合,使得最终每个论元的 embedding 中都含有词语、句子和段落信息,在 PDTB 第 1 层 4 类分类任务上 F 值分别为 41.91%,54.72%,71.54%,34.78%,同时在第 2 层分类任务上取得44.75%的准确率.另外,他们还将该模型用于宾州汉语树库篇章隐式关系的识别,准确率达到 82.56%,与全部标记为最大类别扩展关系的 baseline 相比,提高了 11.63%.Qin 等[48]提出了一个挖掘关系特征的对抗网络来进行隐式关系识别,在 4 类关系分类上取得 46.23%的准确率.Geng 等[49]认为句子结构信息对隐式关系的判定有十分重要的作用,因此应该将句法树信息融入论元的语义编码,他们在将关系论元使用 Bi-LSTM 编码后,将句子的句法树转换成一个二叉树,然后将子节点的信息经过转换后计入父节点信息,最后取得了 62.4%的准确率和 44.2%的 F 值.Wang 等[50]在使用句法树信息之外,也使用了句法树每个节点标签的 embedding,分别在第 1 层和第 2 层语义关系分类中取得了 59.85%和 45.21%的准确率.Dai 等[51]借鉴序列化标注思想,认为句间关系要放在整个篇章中来考察,因此建立了一个篇章级神经网络模型,对显式关系和隐式关系训练不同的分类器,同时在模型最后一层加入了条件随机场(CRF)层,最终取得了 4 分类任务中隐式关系 58.2%的准确率和显式关系 94.46%的准确率.神经网络的应用提高了隐式篇章关系的识别性能,但仍仅有 60%左右的准确率,F 值也不到 50%,仍然无法满足实际应用的需求.

3.4　结构层次树的生成

目前,篇章层次树生成的研究大多基于 RST-DT 展开.Soricut 等[22]使用概率模型构建句级篇章结构树,并在 18 类篇章关系标注上取得 49.0%的 F 值.LeThanh 等[52]分别在句子层面和篇章层面进行篇章结构树的构建,在句子层面使用句法信息和短语信息切分 EDU,以生成句子的篇章结构树,并取得了 66.2%的 F 值.在汉语的篇章关系构建中,张益民等[53]利用主位模式等多个语言学特征,使用向量空间模型对篇章结构进行自动分析.涂眉等[54]先使用序列化标注方法对篇章语义单元进行切分,然后使用最大熵模型对篇章结构进行推导,在清华汉语树库上的实验结果为,当篇章语义结构树高度不超过 6 层时,篇章语义关系标注的 F 值为 63%.可以看到,过去对结构层次树生成的评测主要仍是针对层次生成后的语义关系标注,对结构层次本身的正确与否并无考察.对于含有多个分句的复句或篇章来说,句子之间的层次关系直接反映了它们之间的逻辑语义关系,因此对层次结构树本身的考察是今后研究亟待解决的关键问题之一.

4　研究展望

4.1　主要问题

从上述国内外研究现状可以看出,目前的复句处理研究还存在以下问题:

1) 缺乏一个统一的汉语复句语义分析的理论体系.语言学界对复句关系的划分有多种方法,缺乏一个普遍认同和遵从的标准.因此目前研究使用的复句分类体系划分不同,有的使用两分法,有的使用三分法,有的使用小类分法,有的将英语 PDTB 体系借鉴到汉语中来.无法在同一个平台进行横向比较,不利于汉语复句的进一步研究和建设.

2) 缺乏针对复句的大规模语料库.目前常见的篇

章关系语料库在语料划分粒度上不一致,有的是复句,有的是句群,有的甚至是段落.专门针对复句的语义关系和结构层次划分的语料库还没有.目前仍在建设中的 CAMR 语料库虽然包含了复句间的语义关系和结构层次,但若要作为复句结构语义语料库使用,还必须对关系词、语义关系做更深入细致的描写和标注.

3) 目前国内外的研究主要着眼于篇章语义关系,专门针对复句的研究仍然较少.复句是篇章的组成单位,篇章各层级语段之间存在着高频的复现关系,因此弄清楚复句中各分句的衔接方法和结构层次,篇章关系才能够得到更好地解决.目前的研究中不论是语义关系的判定还是结构层次的划分,都是在篇章层面上进行的,复句相对于篇章来说,篇幅更短,在更短的文本中寻找其语义关系,划分其结构层次,是需要进一步探索的.

4) 国内目前关于复句的研究多是针对二分句的,少部分是针对三分句的研究,而在自然语言中,复句中的分句数目往往更多,其结构层次的复杂程度呈指数级上升,而目前这方面的研究仍然少有涉猎.

5) 隐式语义关系的识别仍然是个难点.虽然隐式语义关系一直是篇章关系研究的热点,近年来神经网络也被应用于隐式语义关系的识别,但由于该任务涉及深层语义理解,难度较大,效果一直不甚理想,目前最好的整体效果也只有 40%~50%,这说明要解决这一难题,仍然需要投入更多的努力.

4.2 未来工作

AMR 在句子语义表示方面有着得天独厚的优势,同时也是下一步篇章语义表示的基础.为了更好地对 AMR 中的句子进行解析,有必要对复句进行更加深入的研究.接下来我们的工作将从以下方面进行:

1) 完善 CAMR 标注体系,制定更符合汉语实际的标注规范,在目前标注的基础上,完善与复句有关的标注内容.

2) 探索多种复句标注体系间的对应关系及转换方法,从而实现复句语义资源的整合利用.

3) 对于有多个逗号隔开的复句,进行论元识别和边界切分.从而为下一步语义关系识别打下基础.

4) 无标记复句的语义关系本身存在模糊性,不同标注者可能对同一无标记复句标注不同的语义关系,对机器来说,这更是一个具有挑战性的问题.因此,应提高标注的内在一致性、寻找方法提高机器自动识别无标记复句语义关系性能.

5) 构建复句逻辑语义结构树,将指代消解、缺省回补等工作与复句逻辑语义结构树结合起来,以更好地对复句语义关系进行抽象表示.

随着自然语言理解中语义分析的深入,AMR 复句解析在信息抽取、自动文摘、机器阅读理解等领域有着重要的研究价值和光明的应用前景,值得不断地研究和探索.

参考文献:

[1] BANARESCU L, BONIAL C, CAI S, et al. Abstract meaning representation for sembanking[C]// Linguistic Annotation Workshop and Interoperability with Discourse. Sofia: Bulgaria, 2013: 178-186.

[2] 曲维光,周俊生,吴晓东,等.自然语言句子抽象语义表示 AMR 研究综述[J].数据采集与处理,2017,32(1):26-36.

[3] 周祖谟.现代汉语讲座[M].北京:知识出版社,1983:154-167.

[4] 胡金柱,舒江波,胡泉,等.复句关系词自动识别中规则的表示方法研究[J].计算机工程与应用,2016,52(1):127-132.

[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] 徐赳赳,WEBSTER J.复句研究与修辞结构理论[J].外语教学与研究,1999(4):16-22.

[7] 黄伯荣,廖序东.现代汉语(增订版)[M].北京:高等教育出版社,2002:123-133.

[8] 邢福义.汉语复句研究[M].北京:商务印书馆,2001:38-47.

[9] 胡明扬,劲松.流水句初探[J].语言教学与研究,1989(4):42-54.

[10] PRASAD R, DINESH N, LEE A, et al. The Penn Discourse TreeBank 2.0[C]// International Conference on the 6th Language Resources and Evaluation. Marrakech, Morocco: LREC, 2008: 2961-2968.

[11] KNIGHT K, BADARAU B, BARANESCU L, et al. Abstract meaning representation (AMR) annotation release 2.0 [DB/OL]. [2017-06-15]. https://catalog.ldc.upenn.edu/LDC2017T10.

[12] 王力.中国语法理论[M].北京:商务印书馆,1951:39-40.

[13] LI B, WEN Y, QU W G, et al. Annotating the little prince with Chinese AMRs[C]// Linguistic Annotation Workshop Held in Conjunction with ACL. Berlin: ACL, 2016: 7-15.

[14] ZHOU Y, XUE N. PDTB-style discourse annotation of

Chinese text[C]// Meeting of the Association for Computational Linguistics: Long Papers. Jeju Island: ACL,2012:69-77.

[15] ZHOU Y, XUE N. The Chinese discourse treebank: a Chinese corpus annotated with discourse relations[J]. Language Resource and Evaluation, 2015, 49(2): 397-431

[16] 周强.汉语句法树库标注体系[J].中文信息学报,2004,18(4):2-9.

[17] 李艳翠.汉语篇章结构表示体系及资源构建研究[D].苏州:苏州大学,2015:65-80.

[18] CARLSON L, MARCU D, OKUROWSKI M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory[J]. Springer Netherlands, 2003, 18(18):2655-2661.

[19] 邢福义,姚双云.汉语复句语料库的建设与利用[C]//第三届 HNC 与语言学研究学术研讨会论文集.北京:北京师范大学出版社,2005:432-437.

[20] 李斌,闻媛,卜丽君,等.英汉《小王子》抽象语义图结构的对比分析[J].中文信息学报,2017,31(1):50-57.

[21] 张牧宇,秦兵,刘挺.中文篇章关系任务分析及语料标注[J].智能计算机与应用,2016,6(5):1-4.

[22] SORICUT R, MARCU D. Sentence level discourse parsing using syntactic and lexical information[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics. Edmonton: ACL,149-156.

[23] HERNAULT H, PRENDINGER H, DUVERLE D A, et al. HILDA: a discourse parser using support vector machine classification[J]. Dialogue and Discourse, 2010, 1(3):1-33.

[24] LIN Z, NG H T, KAN M Y. A PDTB-styled end-to-end discourse parser[J]. Natural Language Engineering, 2014,20(2):151-184.

[25] 洪鹿平.汉语复句关系自动判定研究[D].南京:南京师范大学,2008:13-31.

[26] 胡金柱,俞小娟,李琼,等.基于规则库和聚类分析的复句短语字段的自动识别研究[J].华中师范大学学报(自科版),2008,42(2):190-194.

[27] PITLER E, RAGHUPATHY M, MEHTA H, et al. Easily identifiable discourse relations[C]// International Conference on Computational Linguistics, Posters Proceedings. Manchester: ICCL, 2008:87-90.

[28] PITLER E, NENKOVA A. Using syntax to disambiguate explicit discourse connectives in text[C]// Proceedings of the Association for Computational Linguistics and AFNLP. Singapore: ACL, 2009:13-16.

[29] 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[J].北京大学学报(自然科学版),2014,50(1):118-124.

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

[31] 李艳翠,孙静,周国栋.汉语篇章连接词识别与分类[J].北京大学学报(自然科学版),2015,51(2):307-314.

[32] 杨进才,郭凯凯,沈显君,等.基于贝叶斯模型的复句关系词自动识别与规则挖掘[J].计算机科学,2015,42(7):291-294.

[33] 张牧宇,宋原,秦兵,等.中文篇章级句间语义关系识别[J].中文信息学报,2013,27(6):51-58.

[34] 杨进才,陈忠忠,沈显君,等.二句式非充盈态有标复句关系类别的自动标志[J].计算机应用研究,2017,34(10):2950-2953.

[35] MARCU D, ECHIHABI A. An unsupervised approach to recognizing discourse relations[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics. Philadelphia: ACL, 2002: 368-375.

[36] PITLER E, LOUIS A, NENKOVA A. 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. Singapore: ACL, 2009:683-691.

[37] LIN Z, KAN M Y, NG H T. Recognizing implicit discourse relations in the Penn discourse treebank[C]// Conference on Empirical Methods in Natural Language Processing. Singapore: EMNLP, 2009:343-351.

[38] LOUIS A, JOSHI A, PRASAD R, et al. Using entity features to classify implicit discourse relations[C]// Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics. Tokyo: ACL, 2010:59-62.

[39] RUTHERFORD A T, XUE N. Discovering implicit discourse relations through Brown cluster pair representation and coreference patterns[C]// Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg: ACL, 2014:645-654.

[40] RUTHERFORD A, XUE N. Improving the inference of implicit discourse relations via classifying explicit discourse connectives[C]// Conference of the North American Chapter of the Association for Computational Linguistics. Denvor: ACL, 2015:799-808.

[41] 车婷婷,洪宇,周小佩,等.基于功能连接词的隐式篇章关系推理[J].中文信息学报,2014,28(2):17-27.

[42] 孙静,李艳翠,周国栋,等.汉语隐式篇章关系识别[J].

北京大学学报(自然科学版),2014,50(1):111-117.
[43] 李国臣,张雅星,李茹.基于汉语框架语义网的篇章关系识别[J].中文信息学报,2017,31(6):172-179.
[44] JI Y, EISENSTEIN J. Entity-augmented distributional semantics for discourse relations[J]. Transaction for computational Linguistics(TACL),2014,3:329-344.
[45] ZHANG B,SU J,XIONG D,et al. Shallow convolutional neural network for implicit discourse relation recognition [C]// Conference on Empirical Methods in Natural Language Processing. Lisbon:EMNLP,2015:2230-2235.
[46] LIU Y,LI S. Recognizing implicit discourse relations via repeated reading: neural networks with multi-level attention[C]// Conference on Empirical Methods in Natural Language Processing. Austin: EMNLP, 2016: 1224-1233.
[47] LI H, ZHANG J, ZONG C. Implicit discourse relation recognition for English and Chinese with multiview modeling and effective representation learning[J]. ACM Trans Asian Low-Resour Lang Inf Process, 2017, 16 (3):1-21.
[48] QIN L,ZHANG Z,ZHAO H,et al. Adversarial connective-exploiting networks for implicit discourse relation classification[EB/OL]. [2017-04-01]. http://cn.arxiv.org/abs/1704.00217.
[49] GENG R, JIAN P, ZHANG Y, et al. Implicit discourse relation identification based on tree structure neural network [C] // International Conference on Asian Language Processing. Singapore: IEEE, 2017: 334-337.
[50] WANG Y, LI S, YANG J, et al. Tag-enhanced tree-structured neural networks for implicit discourse relation classification[C] // International Joint Conference on Natural Language Processing. Taipei: AFNLP, 2017: 496-505.
[51] DAI Z, HUANG R. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph[C]// the North American Chapter of the Association for Computational Linguistics. New Orleans: ACL, 2018, June 1-6.
[52] LETHANH H, ABEYSINGHE G, HUYCK C. Generating discourse structures for written texts[C]// Proceedings of International Conference on Computational Linguistics. Association for Computational Linguistics. London: ACL, 2004: 329-335.
[53] 张益民,陆汝占.一种混合型的汉语篇章结构自动分析方法[J].软件学报,2000,11(11):1527-1533.
[54] 涂眉,周玉,宗成庆.基于最大熵的汉语篇章结构自动分析方法[J].北京大学学报(自然科学版),2014,50(1):125-132.

A Survey on the Study of Compound Sentences with Chinese Abstract Meaning Representation

WEI Tingxin[1,2], QU Weiguang[2,3,4*], SONG Li[2], DAI Rubing[2]

(1. International College for Chinese Studies, 2. School of Chinese Language and Literature, 3. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210097, China; 4. Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou 350121, China)

Abstract: Abstract meaning representation (AMR) is a novel framework of representing sentential meaning. Due to linguistic characteristics of the Chinese language, Chinese AMR(CAMR) annotates the semantic compound-sentence meaning, which is ignored in English AMR annotation. Sentence is the elementary unit of CAMR and hierarchical tree structure is used to represent the logical relations of all minimal sentences in compound sentences. As arguments can be shared, the graph based on the tree structure expresses the semantic meaning of compound sentences more comprehensively. This paper introduces the current resource construction and methodology of compound sentence relations and discourse relations, as well as points out the key problems lying in present studies. Then future work is discussed.

Key words: Chinese abstract meaning representation(CAMR); compound sentence; discourse relation