文章编号:1000-5862(2015)02-0124-08

中文篇章级句间关系自动分析

姬建辉,张牧宇,秦 兵*,刘 挺
(哈尔滨工业大学计算机学院,黑龙江 哈尔滨 150001)

摘要:篇章级句间关系分析包括语义单元的切分和各个单元之间的语义关系识别.已有的研究主要面向英文,到目前为止,尚无可用的中文篇章级句间关系自动分析系统发布.在中文篇章关系语料库的基础上,首次实现面向中文的篇章级句间关系自动分析系统,包括语义单元切分、连词识别、显式语义关系识别以及隐式语义关系识别等.实验结果显示:该系统在显式句间关系识别上 F-score 为 89.8%,隐式句间关系识别上 F-score 为 55.5%.

关键词:中文篇章分析;篇章句间关系;语义单元
中图分类号:TP 391 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2015.02.03

0 引言

随着底层自然语言处理技术的日益成熟,篇章级语义分析逐渐成为研究热点.作为篇章语义分析的重要内容之一,篇章级句间关系分析(Discourse relation analysis)也开始受到研究人员的关注.该研究是在词汇分析、短语分析的基础上,对篇章进行深层次语义分析,在文本单元分割的基础上,判断各个单元(分句、复句、句群)之间的内在逻辑关系(例如:因果),从而理解整个篇章的组成架构.该研究为情感分析[1]、文本连贯性[2]等 NLP 任务在篇章理解层面提供了有力的技术支持.

篇章级句间关系分析通常包括 2 个部分内容:(i)基本语义单元(Elementary discourse unit,EDU)[3]识别;(ii)语义单元之间语义关系的识别.所谓基本语义单元是指句子中具有独立性和表述性的最小的语法单位,它表达了一个基本的、完整的语义.多个 EDU 通过一定的组织方式连接成更为复杂的语义单元如句子、段落,以至于完整的篇章.根据 2 个语义单元之间是否通过显式的连词连接,语义单元之间的句间关系又可以进一步分为显式篇章句间关系(Explicit discourse relation,简称显式关系)与隐式篇章句间关系(Implicit discourse relation,简称隐式关系)[4].

例 1 [懒惰是人的天性]$_{edu1}$,[但是]$_{Connective}$[我们要想成功,一定要多多努力]$_{edu2}$!

例 2 [昨天下了一天的雨]$_{edu1}$,[晾在外面的衣服全都淋湿了]$_{edu2}$.

例 1 中 2 个语义单元表达了一种转折关系,通过关联词"但是"连接起来,称之为转折类型的显式关系.而例 2 中 2 个语义单元之间没有关联词出现,而是通过逻辑上的因果关系连接在一起,称之为因果类型的的隐式句间关系.

目前已有的研究主要面向英文,在人工标注的篇章关系语料基础上展开,探索了句间语义关系分析的各个子问题.此类研究大多基于 Penn Discourse Tree Bank (PDTB)[4],也是目前可用的规模最大的篇章关系语料库,该语料库以连词为核心标注了连词连接的语义单元以及 4 大类句间关系,分别为 Comparison(比较关系)、Expansion(扩展关系)、Contingency(因果关系)和 Temporal(时序关系).在中文方面,文献[5]首次探索了中文篇章关联词的标注,并分析了中文篇章级句间关系的特点.文献[6-7]讨论了中文篇章关系分析中的逗号问题,并基于该研究提出中文 EDU 的切分方案.此外,文献[8]讨论了面向中文的篇章关系语料库的标注问题,并尝试进行了中文篇章句间关系的识别[9].据笔者所

收稿日期:2015-01-17
基金项目:国家自然科学基金(61133012,61273321)和国家"863"前沿技术研究(2012AA011102)资助项目.
通信作者:秦 兵(1968-),女,黑龙江哈尔滨人,教授,博士生导师,主要从事自然语言处理、文体、挖掘和情感分析等方面的研究.

知,目前在中文篇章级句间关系分析方面,依然缺乏可用的篇章自动分析系统,这也限制了篇章级句间关系分析在中文自然语言处理上的实际应用.

本文提出首个中文篇章级句间关系分析系统,针对该问题的各个子任务分别提出相应的解决方案,实现了面向中文的篇章级句间关系自动分析. 在哈尔滨工业大学信息检索研究中心开发的中文篇章关系语料库 HIT-CDTB [3] 的基础上,分别探索了篇章基本语义单元的自动切分、篇章关联词识别、显式句间关系识别以及隐式句间关系识别等任务. 针对一篇输入的原始文章,模型按照如下步骤展开:

(i) 对文章进行基本的底层 NLP 预处理;

(ii) 针对每个句子,在对其进行短语结构分析的基础上,实现了基于递归的基本语义单元自动切分,将句子分割成具有独立性和表述性的语义单元集合 EDU;

(iii) 针对语义单元之间连接方式,本文在构建中文连词词典的基础上开发了基于 SVM 的中文连词识别模型,通过该模型来识别句子中连接语义单元的连词;

(iv) 根据语义单元之间的连接是否依靠候选连词,将语义单元之间的语义关系识别划分为通过连词连接的显式句间关系识别和通过语义关系连接的隐式句间关系识别.

针对显式句间关系,本文直接使用基于统计规则的方法进行处理,并取得了较好的效果. 针对隐式句间关系,在中文篇章级句间关系语料的基础上,实现了基于 SVM 的有指导关系识别模型,从而实现了第 1 个面向中文的篇章级句间关系端到端分析系统. 实验结果显示: 句内语义单元切分模块的准确率达到了 54.6%,篇章关联词识别 F 值达到 87%,显式句间关系识别的准确率达到了 89.8%,隐式句间关系识别的准确率为 55.5%.

1　篇章级句间关系自动分析系统

句子是由实体、事件等信息按照一定的语义关系连接组成. 篇章级句间关系分析是在词汇级、句子级之上的深层次篇章结构理解. 本文构造的篇章自动分析系统包括 4 个模块: 基本语义单元切分模块、候选连词识别模块、显式句间关系识别模块以及隐式句间关系识别模块. 系统的框架如图 1 所示.

1.1　语义单元切分

语义单元是指在一篇文章中具有独立性和表述性的最小的语法单位. 语义单元通过不同的连接方式组成分句、句子,进而组成段落以及篇章. 因此,篇章分析的首要任务就是对篇章中的句子进行基本语义单元的切分. 基本语义单元的自动切分一直是篇章分析中的一大挑战. 在英文方面,现有的研究大部分是在概率模型的基础上对句子进行语义单元的切分,而在中文方面,因为中文的逗号特色问题,现有的研究集中在针对逗号进行消歧来将中文长句子切分为短句子,但是对于句子内部的语义单元切分却没有进行探索.

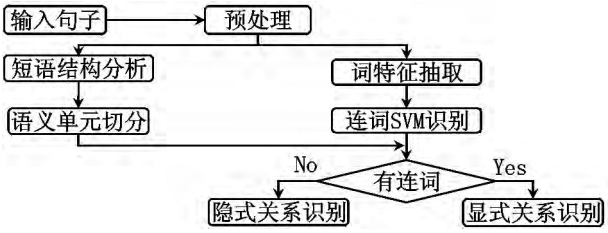

图 1　中文篇章级句间关系分析结构图

一个句子是一棵由语言单元组成的树,语义单元之间存在层次关系. 在例 1 中共存在 2 个句间关系,一个是由“但是”连接的显式的转折关系,如例 3 所示. 其中的 edu2 中还存在一个隐式的条件关系,如例 4 所示.

例 3　[懒惰是人的天性]$_{edu1}$, [但是]$_{Connective}$ [我们要想成功,一定要多多努力]$_{edu2}$!

例 4　懒惰是人的天性,但是 [我们要想成功]$_{edu1}$, [一定要多多努力]$_{edu2}$!

例 5　[懒惰是人的天性]$_{edu1}$, [但是]$_{Connective}$ [我们要想成功]$_{edu2}$, [一定要多多努力]$_{edu3}$!

假设仅仅使用逗号来进行语义单元分割,那么例 1 中将会形成 3 个同级别的语义单元,如例 5 所示. 从中可以看出“但是”表达的转折关系连接的 2 个语义单元是 edu1 和 edu2,并没有包含 edu3,错误的主要原因在于仅仅使用逗号进行分割并没有考虑到句子内部的语义单元是有层次的,正如篇章中的句子也是按照段落来组织表达一种层次的.

在篇章中可能出现的信息,比如实体、事件等都是无穷尽的,但是篇章中的信息却是按照统一的习惯规则进行组织连接的. 因此 ,在对句子进行短语结构分析的基础上,提出了基于启发式规则的基本语义单元切分方法. 例 1 的短语结构树如图 2 所示.

在一个句子中重要的词是其中的动词和名词,其中动词在依存分析中称之为核心词 Head. 而语义单元是句子中具有独立性和表述性的最小单元,因此最基本的语义单元至少包含一个动词. 为此,在短语结构的基础上,以动词为核心构建基本语义单元的识别规则. 值得注意的是,篇章中存在多个层次的语义单元,并且可能以嵌套的形式出现,因此提出基于递归的语义单元识别规则如下: (i) 基本语义单元

为一个 VP(动词短语);(ii) 一个 IP(简单从句) 作为一个语义单元当且仅当该 IP 包含一个语义单元;(iii) 语义单元可以由基本语义单元通过组合而成;(iv) 每个句子至少由一个语义单元组成;(v) 句间关系存在于相同层次的语义单元之间.

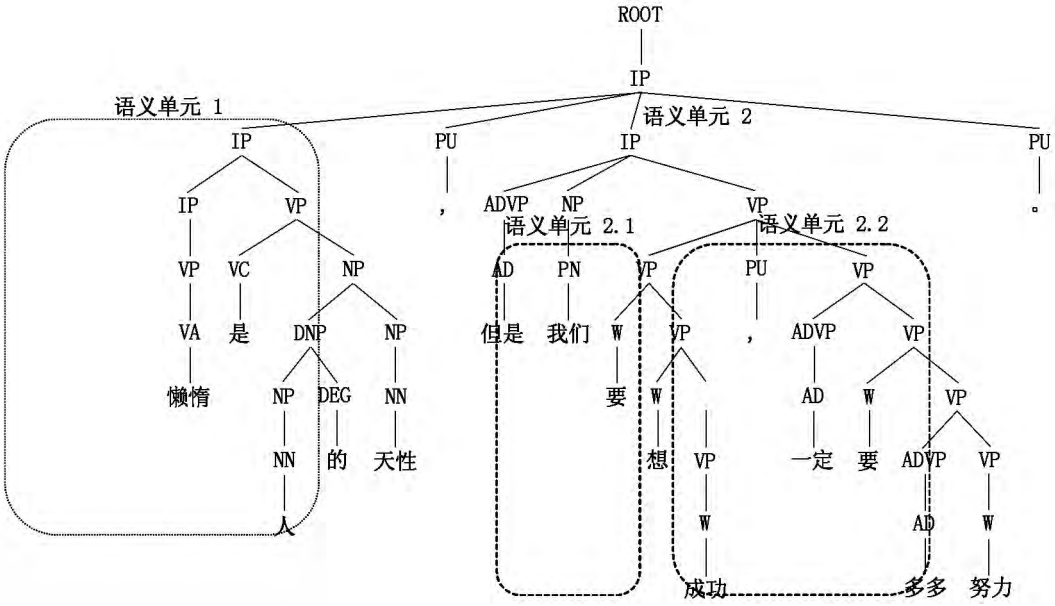

图 2 基于短语结构分析的语义单元树

最终篇章中的每一个句子就表示为一棵由基本语义单元组成的语义单元树,最终形成的语义单元树如图 2 所示. 整个句子作为一个语义单元由语义单元 1 和语义单元 2 连接而成. 而语义单元 2 由语义单元 2.1 和语义单元 2.2 组成. 通过语义单元的树形表示,有效地解决了逗号分隔语义单元所不能表示的语义单元之间的层次化.

需要指出的是,以 VP 为核心的识别方案虽然拥有较好的性能,但并不能覆盖所有的实例,尤其是在中文中更是如此. 这主要是由于中文存在一定程度的随意性,很多时候名词短语(NP) 等同样可以具有相对独立的语义. 考虑到语义单元切分任务本身的复杂性,本文并没有尝试覆盖全部实例.

1.2 连词识别

篇章中基本语义单元之间的连接方式有 2 种:(i) 语义单元之间通过实际存在的语法单位即连词连接;(ii) 语义单元之间通过内在的逻辑语义关系连接. 为了计算 2 种方式的比例,首先在中文篇章关系语料库中计算中文连词的使用频率,统计结果显示在 1 096 篇语料中,总共出现了 20 184 个句子,其中包含连词的句子数目为 11 574 个,比例为 57.3%,结果表明在实际语料中人们为了表达一个完整的语义常使用连词来连接 2 个语义单元. 为了进行篇章自动分析,在对输入的句子进行基本语义单元切分之后,需要判断句子中是否存在连接语义单元的候选连词.

1.2.1 中文连词词典 为了进行连词识别,首先在人工标注语料的基础上构建了中文连词词典. 该词典中共包含了 604 个连词,其中连词分为 2 类. 第 1 类是单个词连接 2 个语义单元,比如: 例 1 中的“但是”,类似的单个连词有“因此”、“但”等. 第 2 类连词是通过使用 2 个词来连接 2 个语义单元,比如:“虽然司机接到了无线电通知,但是根本来不及刹车.”中的“虽然…但是”2 个词组合使用表达了转折的语义,类似的组合连词有“只有…才”、“如果…那么”等. 中文中常见连词的出现次数分布规律如表 1 所示.

表 1 按照出现次数区间划分显式关联词

出现次数 R	关联词个数	占总个数比例/%	出现次数	占总次数比例/%
$50 < R$	40	2.72	7 079	61.45
$10 < R \leqslant 50$	101	6.86	2 214	19.22
$2 \leqslant R \leqslant 10$	342	23.23	1 237	10.74
$R = 1$	989	67.19	989	8.59
总数	1 472	100.00	11 519	100.00

从数据中得出出现次数大于 10 次的连词占据了连词种类的 9.58%,但是他们在语料中出现的总

次数占据了总次数的 80.67%,数据表明在实际语料中显式关联词的使用比较集中. 同时为了计算单个连词和组合连词的分布,在连词词典的基础上计算得到: 在所有的连词中一个连词由单个词组成的比例为 87.04%,而由多个词组合而成的比例为 12.96%. 以上数据表明,在中文语料中,连词的使用情况比较集中,中文连词词典的构建和分析为后续的连词识别工作提供了坚实的数据基础.

1.2.2　连词自动识别　一个词既有可能在句子中充当连词的角色,也有可能在句子中充当非连词的角色. 为了验证中文连词词典的准确性,首先在语料库中计算连词词典中的词在语料中作为连词和不作为连词的次数分布,即连词词典中的词在语料中作为连词出现的可能性. 在中文篇章句间关系语料库的基础上进行了第 2 次标注,主要是标注那些在中文连词词典中出现,但是在某些句子中并没有充当连词角色的词. 得到的结果如表 2 所示.

表 2　一个词作为连词出现的次数在该词出现的总次数的比例分布

一个词作为连词的次数/该词总次数	在词典中这样的词占据的比例/%
1.0	71.2
0.9 ~ 1.0	14.4
0.8 ~ 0.9	5.9
0.7 ~ 0.8	2.7
0.6 ~ 0.7	0.1
0.0 ~ 0.6	5.6

从表 2 可以看出,在连词词典中大部分的词一旦出现就是作为连词出现的,但是仍然有一部分的词是有歧义性的. 即: 虽然该词在连词词典中出现,但是在有些情况下该词是不充当连词角色的. 为此在中文连词词典的基础上提出了基于 SVM 分类的连词识别方法. 对句子中的每个词判断其是否是一个连词是一个典型的 2 元分类问题. 在语料库的基础上抽取人工标注的连词的上下文特征作为样本训练了基于 SVM 分类的连词识别模型. 抽取连词的特征模板如表 3 所示.

表 3　连词识别特征模板

特征	取值	详细说明
Length	0 ~ 5	该词的长度
posInLine	0 ~ 300	该词在句中的位置
ambiguity	0 ~ 1	该词在语料中作为连词和不作为连词的比例
occurInDict	0 ~ 1 000	该词在连词词典中出现次数
pos	'n' - 'wh'	该词在句中的词性
prevPos	'n' - 'wh'	该词在句中前一个词的词性
nextPos	'n' - 'wh'	该词在句中下一个词的词性

1.3　句间关系识别模块

篇章自动分析的第 3 步是分析语义单元之间的句间关系. 根据语义单元之间是否有显式的连词连接,句间关系分为有连词连接的显式句间关系和无连词连接的隐式句间关系. 笔者首先计算了在语料中的显式句间关系和隐式句间关系的比例,结果表明在人工标注的所有句间关系中,显式句间关系占据的比例为 53.91%; 隐式句间关系占据的比例为 46.09%. 可以看出,在实际的语料中显式关系和隐式关系占据的比例相差不大.

1.3.1　显式句间关系识别　显式句间关系是指由连词连接的 2 个语义单元之间的句间关系. 因此,在显式句间关系中,连词是一个非常重要的特征. 目前英文上面关于显式句间关系的识别主要采用的是依靠连词来识别 2 个语义单元之间的句间关系. 为了验证在中文方面使用连词来识别句间关系的可能性,笔者在前文构建的连词词典的基础上计算了人工标注中的每个连词表达句间关系种类的可能性分布,即一个连词指示多个句间关系的概率分布. 具体的计算方法为

$$P(c_i, s_j) = Num(c_i, s_j) \Big/ \sum_{s_j \in S} Num(c_i, s_j),$$

其中 c_i 对应某一关联词; s_j 为待计算的关系类型; S 为所有关系类型的集合. 结果如表 4 所示. 从表 4 可以看出,超过 91% 的中文连词指示的显式句间关系不超过 2 类,因此在中文中可以使用连词来识别显式句间关系.

表 4　连词指向关系的歧义性分布

一个连词指示的关系种类	出现次数	占据比例/%
连词只指示 1 个关系	5 225	66.6
连词共指示 2 个关系	1 997	25.4
连词共指示 3 个关系	594	7.6
连词共指示 4 个关系	32	0.4
连词指示超过 5 个关系	0	0

1.3.2　隐式句间关系识别　隐式句间关系是指语义单元之间没有通过显式的连词连接,而是通过逻辑上的语义关系连接在一起. 在人工标注的时候也是由标注人员通过判断 2 个语义单元之间是否存在逻辑上的语义关系来标注隐式句间关系. 因此隐式句间关系的识别不同于显式句间关系的识别.

隐式句间关系的识别本质上是一个分类问题,即判断 2 个语义单元之间逻辑语义单元之间的关系类别. 为了进行自动识别隐式句间关系,采用有指导的方法在训练语料上抽取特征训练了隐式句间关系识别模型. 抽取的特征如下所示:

1) 情感极性: 不同的极性信息常常指示特定的篇章句间关系类型,如下例

例 6　这地方比较 [脏乱],没想到他们觉得很

[幸福].

例 6 中,"脏乱"指示"贬义"的极性信息,而"幸福"指示"褒义",两者的极性信息相反,同时指示了 2 个分句间的转折关系. 基于以上分析,本文引入极性特征. 利用大连理工情感词典分析语义单元中的词汇,获取对应的情感极性,并将最后的结果作为语义单元的情感极性特征. 这里用 1 指示褒义,用 0 表示贬义.

2) 关键词特征: 由于语料限制,并不能完全枚举所有可能出现的连词,仍有部分连词没有被收录在构建的中文连词词典中. 如下例

例 7 [这次毕业 10 年聚会很多人都来了],[唯独肖大宝没有来参加].

Kc03A01 = 但是,只是,可是,不过,然而,可然,而,但,然则,唯独,而是(同义词词林)

例 7 中,"唯独"明确指示了例外关系. 该词汇并没有出现在关联词词典中,却被同义词词林所覆盖,并且和连词词典中的某些词汇位于同一类别,称类似于"唯独"的词汇为"关键词". 这提示大家,如果利用同义词词林中的类别信息,有助于挖掘更多未被覆盖的指示词信息. 考虑到同义词词林中,相同类别的词均有一个统一的类别编号,因此利用关联词词典获取相应的类别编号,并检查待分析的句子中是否有同类别的词汇出现,并使用该类别标号作为特征使用.

3) 核心动词对: 作为句子的主要成分,动词往往在语义表达中起很重要的作用,动词之间的关系常常反映了句子间的语义关系. 如下例

例 8 前些天下雨的时候隔壁的王阿姨摔倒了,到现在都住院快半个月了.

例 8 中,"摔倒—住院"代表了一种隐式的因果关系,同时也指示了 2 个分句之间的因果关系. 因此在识别隐式句间关系时,通过挖掘动词之间的搭配特性,有助于更好地识别篇章句间关系类型. 为了减少数据的稀疏性,同样将抽取到的动词在同义词词林中向上进行泛化,获取动词所属的类别,并将泛化结果配对构成核心动词对特征.

对于情感极性特征,分别使用 1、0 对应褒义和贬义; 对于关键词特征和动词次对特征,直接使用同义词词林中的数字编号作为特征表示. 随后,采用以上特征集合训练 SVM 分类器,对不同的隐式句间关系进行识别.

2 实验设计与分析

2.1 实验语料

本文使用哈尔滨工业大学社会计算与信息检索研究中心开发的中文篇章句间关系语料库[6],包含从 Ontonotes4.0 中筛选的 1 096 篇文章,包含了 5 个领域,包括 bc(broad conversion)、bn(broad news)、mz(magazine)、nw(news) 和 wb(web). 每篇文章都人工标注了其中的语义单元、链接语义单元连词以及语义单元之间的显式句间关系或者隐式句间关系. 从中选取了 4 个领域的 535 篇文章作为实验用的语料,主要是剔除了 bc 谈话领域内的文章以及其它领域内句间关系少于 10 篇文章.

2.2 实验设置与结果分析

2.2.1 语义单元切分模块 根据语义单元的定义,编写了基于递归的句子级别的语义单元切分伪代码. 按照物理顺序将一个句间关系涉及到的 2 个语义单元分别称之为 EDU1 和 EDU2. 同时为了缩减语义单元树的深度,对叶子节点的基本语义单元进行向上单链泛化,即如果在语义单元树中父亲节点只有一个孩子叶节点,那么将该孩子叶节点从语义单元树中去掉.

为了验证语义单元切分的准确率,随机从语料中抽取了 219 个人工标注的句间关系,共 219 × 2 个语义单元,并将程序自动抽取的语义单元和人工标注的语义单元进行了对比. 统计结果显示在所有的 438 个语义单元中,2 个 EDU 完全相同的个数为 239,准确率达到了 54.6%. 在结果分析发现准确率主要在于人工标注的语义单元边界不太统一造成.

算法 1 语义单元切分.

输入: 原始无结构句子;

输出: 该句子对应的语义单元树的根节点;

(i) 对原始句子进行短语结构分析得到短语结构树的根节点 Root;

(ii) 针对一个 Root,递归分析各个孩子节点 root;

(iii) 针对根节点的每个孩子节点 for each child in root->children;

a) 若该节点包含一个 VP,则该节点具有 EDU(基本语义单元)属性;

b) 若该节点包含一个或者多个具有 EDU 属性的孩子节点,则该节点有 EDU 属性;

c) 该节点的下面 EDU 属性节点个数 = 具有 EDU 属性的孩子节点的个数;

d) 若该节点仅有一个 EDU 属性的孩子节点,孩子节点的 EDU 属性转移到父节点.

2.2.2 连词识别模块 首先使用是否出现在中文连词词典中来判断一个词是否是连词作为 Base-Line,得到的实验数据为如表 2 所示. 其中根据连词出现次数进行过滤构建了不同大小的词典. 从表 2

可以看出,基于词典的连词识别效果不是太理想,主要原因是一个句子是由多个词组成的,而仅仅依靠词典来识别其中的候选连词会出现一个句子中多个词都被标示为候选连词,因此造成了连词识别的准确率和召回率效果不高. 而且由前面的分析可知连词词典中前 9% 的连词出现的次数占据了 80%. 因此连词词典的大小对连词识别的影响效果不大.

随后,根据前面定义的连词特征模板,在1 096 篇语料中共抽取了 40 329 个样本,其中正样本为 11 192个,占据比例为 27.75%. 在此基础上随机抽取了 32 260 个样本作为训练样本,8 069 个样本作为测试样本,得到的连词识别效果如表 5 所示. 从表 5 可以看出,使用基于 SVM 的连词识别模型有效地提高了连词识别的准确率和召回率.

表 5　基于词典和 SVM 分类模型的连词识别效果

实验	训练	测试	*P*	*R*	*F*
基于词典(600 个连词)	–	40 329	0.71	0.67	0.69
基于词典(260 个连词)	–	40 329	0.70	0.65	0.68
基于词典(121 个连词)	–	40 329	0.71	0.63	0.67
基于 SVM 分类	32 260	8 069	0.86	0.87	0.87

2.2.3　**显式句间关系识别**　为了验证使用连词来识别句间关系的准确性,实验采用的语料为中文篇章语料里面的 525 篇句间关系语料作为测试语料. 在实验中使用连词指示概率最大的句间关系来预测连词所连接的 2 个语义单元之间的句间关系,得到的数据结果如表 6 所示.

表 6　以中文连词词典为基础的显式句间关系的识别

关系	标注	识别	识别正确	*P*	*R*	*F*
1 时序关系	502	525	490	0.93	0.98	0.95
2 因果关系	1 571	1 614	1 554	0.96	0.99	0.98
3 条件关系	685	682	664	0.97	0.97	0.97
4 比较关系	2 260	2 112	2 045	0.97	0.90	0.94
5 扩展关系	2 208	2 571	1 996	0.78	0.90	0.84
6 并列关系	765	509	397	0.78	0.52	0.62

从表 6 可以得出 6 类显式句间关系识别准确率的宏平均为 0.898. 因此,在中文显式句间关系识别方面,仅仅使用连词词典就可以达到较好的识别效果. 但是注意到前 4 种显式句间关系使用连词来识别效果良好,但是后 2 种显式句间关系的效果不理想. 通过对连词词典分析以及观察语料还发现,指示扩展关系或者并列关系所属的语义单元之间语义关联度没有前 4 种关系那么强,造成了连接语义单元的连词的歧义性比较大. 但是总体来看,6 类关系识别的准确率的宏平均为 0.89,召回率的宏平均为 0.87,而 F 值的宏平均为 0.88,因此仅仅使用连词来识别中文显式句间关系是可行的.

2.2.4　**隐式句间关系识别**　根据特征抽取方法,在篇章关系语料库中共抽取隐式句间关系样本数目为 7 169 个. 在语料库中标注的隐式句间关系中 1 时序关系 3 条件关系占据的比例太少,因此本文并没有抽取 1 时序关系和 3 条件关系的训练样本. 最终基于 SVM 训练的隐式句间关系识别模型的效果为如表 7 所示. 从表 7 中可以看出,隐式关系中 4 比较关系和 2 因果关系的识别效果不理想,主要原因在于它们在隐式关系中出现的比例太少.

表 7　基于 SVM 分类的隐式句间关系识别

关系	样本数	占据比例/%	*P*	*R*	*F*
2 因果关系	1 182	16.31	0.59	0.06	0.11
4 比较关系	472	6.60	0.33	0.01	0.02
5 扩展关系	3 963	55.48	0.65	0.93	0.77
6 并列关系	1 333	18.61	0.65	0.54	0.59

3　相关工作

作为篇章分析的基础,英文方面已经人工标注了大量的篇章句间关系语料库. Williamand Thompson [10] 提出了基于修辞结构理论的树库. Miltsakaki 等 [4] 公布了宾州篇章关系树库 PDTB(Penn Discourse Tree Bank),该树库标注了篇章中出现的连词以及连词连接的 2 个语义单元范围,并标注了语义单元之间存在的显式篇章句间关系和隐式篇章句间关系. Xue Ninewen 等 [5] 对中文连词进行了分类并介绍了中文篇章关系语料库中连词的标注. 张牧宇等 [8] 提出了面向中文的篇章关系体系并标注了中文篇章关系语料库. Oza 等 [11] 介绍了印度语篇章关系树库的构建.

在基本语义单元识别方面,Soricut 等 [3] 通过判断句子中的每个词作为语义单元边界的概率实现了基于概率模型的篇章语义单元自动切分模型并构建了句子级别的篇章分析树,该方法的实验结果相比传统的基于决策树的方法错误减少了 18.8%. B. Wellne [12] 避开确定显式连词连接的 2 个语义单元的准确范围,转而去抽取连词连接的 2 个语义单元的核心词. Rajen Subba 等 [13] 在 RST-DT 语料基础上通过使用词性特征、语法特征、关键词特征以及标点符号等信息训练了基于神经网络的篇章语义单元切分模型,该模型的 F 值达到了 84.41%. B. Elwell [12] 将连词分为从属连词、并列连词和关联副词并针对不同类型的连词训练了不同的语义单元范围识别模型. Milan Tofiloski [14] 通过使用词法和语法等信息并结合一些列规则开发了基于启发式规则的篇章语义

单元切分模型(SLSeg).在中文篇章语义单元处理方面,Jin Meixun[15-16]抽取中文逗号出现的上下文信息对逗号进行分类,分类的准确度达到了87.1%,并根据逗号的分类结果将中文长句子切分.Xue Mianwen[6]对中文逗号进行了消歧研究.Yang Yaqi[7]通过对逗号消歧完成了中文语义单元的切分.

篇章关系识别方面,Daniel Marcu[9]通过使用模板抽取语料完成了无监督的篇章关系分类方法.Emily Pitler[17]在PDTB语料的基础上分析了连词的歧义性和依靠连词来识别显式篇章关系的效果,实验结果表明使用连词来识别显式句间关系可以达到93.09%的准确率.随后,Lin Zhiheng等[18]在PDTB语料上使用语义单元的上下文特征、词对信息以及依存特征来识别隐式篇章关系,并获得了40.2%的准确率.Annie Louis[19]通过使用实体特征来识别隐式关系,但是不如使用语法特征的自动识别效果.Wang Xun等[20]使用聚类的方法选择合适的隐式篇章关系训练语料,有效地提升了隐式篇章关系的识别效果.D. Souza等[21]通过引入语言学知识提高了隐式篇章关系下的时序关系的识别效果.在中文篇章关系识别方面,Huang Hen-Hsen等[22]通过手动标注语料完成了基于中文的篇章关系分析.张牧宇[23]在构建中文篇章关系语料库的基础上完成了面向中文的篇章级句间语义关系的识别.

在篇章语义自动识别方面,Rajen Subba[24]提出了一个基于归纳逻辑编程的篇章分析系统,该系统除了使用传统的语法信息外还从wordNet中学习规则并使用了文本之间的相似性来完成篇章的自动分析.H. Hernault[25]在RST-DT语料的基础上开发了基于SVM分类模型的篇章分析系统(High-Level Discourse Analyzer).Lin[18]在PDTB语料库的基础上实现了第1个端到端的英文篇章自动分析系统.

4 结论与展望

本文首次在中文篇章关系语料库的基础上完成了面向中文的篇章自动分析,其中包括了中文语义单元切分、连词识别以及语义单元之间的显式句间关系和隐式句间关系的识别.得到以下结论:

1)在语义单元方面,提出了依靠短语结构分析来完成句子的语义单元切分,达到了很好的效果;

2)对于连词识别,首次构建了中文连词词典并实现了基于SVM分类的中文连词的自动识别模型;

3)在显式句间关系和隐式句间关系方面,借鉴了英文方面比较成熟的方法,并根据中文特有的特点开发了对应的句间关系识别方法;

4)在此基础上开发了一个端到端的在线展示篇章级句间关系分析网站,通过篇章的自动分析,为后续的自然语言处理任务提供了很好的篇章级别的支持;

5)在后续的过程中可以通过不断地更新篇章自动分析的各个模块来提高整个篇章自动分析效果.

5 参考文献

[1] Wang Fei,Wu Yunfang,Qiu Likun. 2012. Exploiting discourse relations for sentiment analysis [EB/OL]. [2014-11-19]. http://www.aclweb.org/anthology/C/C12/C12-2128.pdf.

[2] Lin Ziheng,HweeTou Ng,Min-Yen Kan. Automatically evaluating text coherence using discourse relations [EB/OL]. [2014-11-19]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.207.6356.

[3] Soricut,Radu,Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information [EB/OL]. [2014-12-23]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.4284.

[4] Miltsakaki,Eleni,Rashmi Prasad,et al. The penn discourse Treebank [EB/OL]. [2014-11-20]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.9607&rep=rep1&type=pdf.

[5] Xue Nianwen. Annotating discourse connectives in the Chinese Treebank [EB/OL]. [2014-12-21]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.139.6457.

[6] Xue Nianwen,Yang Yaqin. Chinese sentence segmentation as comma classification [EB/OL]. [2014-12-23]. http://dl.acm.org/citation.cfm?id=2002859.

[7] Yang Yaqin,Xue Nianwen. Chinese comma disambiguation for discourse analysis [EB/OL]. [2014-12-23]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.367.285.

[8] 张牧宇,秦兵,刘挺. 中文篇章级句间语义关系体系及标注[J]. 中文信息学报,2014,28(2):28-35.

[9] Daniel Marcu,Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations [EB/OL]. [2014-12-23]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.669.

[10] Mann,William C,Sandra A Thompson. Rhetorical structure theory: toward a functional theory of text organization[J]. Text,1988,8(3): 243-281.

[11] Oza, Umangi. The Hindi discourse relation bank [EB/OL]. [2014-12-23]. http://dl.acm.org/ft_gateway.cfm?id=1698410&type=pdf&CFID=482731916&CFTOKEN=24330680.

[12] Wellner B, Pustejovsky J. Automatically identifying the arguments of discourse connectives [EB/OL]. [2014-12-23]. http://www.aclweb.org/anthology/D07-1010.

[13] Rajen Subba, Barbara Di Eugenio. Automatic discourse segmentation using neural networks [EB/OL]. [2014-12-23]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.129.3515.

[14] Milan Tofiloski, Julian Brooke, Maite Taboada. A syntactic and lexical-based discourse segmenter [EB/OL]. [2014-12-23]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.7427.

[15] Jin Meixun, Kim Miyoung, Kim dongil, et al. 2004. Segmentation of Chinese long sentences using commas [EB/OL]. [2014-12-23]. http://truth.yust.edu/dongil/research/inter-13.pdf.

[16] Elwell Robert, Jason Baldridge. Discourse connective argument identification with connective specific rankers [EB/OL]. [2014-12-23]. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4597192&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Ficp.jsp%3Farnumber%3D4597192.

[17] Emily Pitler, Mridhula Raghupathy, Hena Mehta, et al. Easily identifiable discourse relations [EB/OL]. [2014-12-23]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.324.4486&rep=rep1&type=pdf.

[18] Lin Ziheng, Minyen Kan, Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank [EB/OL]. [2014-12-24]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.150.9118.

[19] Annie Louis, Aravind Joshi, Rashmi Prasad, et al. Using entity features to classify implicit discourse relations [EB/OL]. [2014-12-24]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.187.4885.

[20] Wang Xun, Li Sujian, Li Jiwei, et al. Implicit discourse relation recognition by selecting typical training examples [EB/OL]. [2014-12-24]. http://wing.comp.nus.edu.sg/~antho/C/C12/C12-1168.pdf.

[21] D'Souza, Jennifer, Vincent Ng. Classifying temporal relations with rich linguistic knowledge [EB/OL]. [2014-12-24]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.378.1441.

[22] Huang HenHsen, Chen Hsin-Hsi. Chinese discourse relation recognition [EB/OL]. [2014-12-24]. http://aclweb.org/anthology/I/I11/I11-1170.pdf.

[23] **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): 51-57.

[24] Subba, Rajen, Barbara Di Eugenio. An effective discourse parser that uses rich linguistic information [EB/OL]. [2014-12-24]. http://dl.acm.org/citation.cfm?id=1620837.

[25] Hernault H, Prendinger H, Duverle D A, et al. HILDA: a discourse parser using support vector machine classification [J]. Dialogue & Discourse 2010, 1(3): 1-33.

The Chinese Discourse Parser

JI Jianhui, ZHANG Muyu, QIN Bing*, LIU Ting

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin Heilongjiang 150001, China)

Abstract: Discourse relation analysis usually focuses on identifying elementary discourse units (EDUs) and recognizing the discourse relation between EDUs. Previous work on this topic primarily focused on English. To the best of our knowledge, there is no publicly available tool for Chinese discourse relation analysis. In the work, based on the human-annotated corpus, the first end-to-end discourse relation analysis system for Chinese, which includes elementary unit segmentation, discourse connective disambiguation, explicit relation recognition, and implicit relation recognition has been developed. In the experiments, the model achieves a F-score of 89.8% in explicit relation recognition and 55.5% in implicit relation recognition.

Key words: Chinese discourse parser; discourse relation; elementary discourse unit

(责任编辑: 冉小晓)