《中文信息学报》稿件排版格式

文章编号：1003-0077（2011）00-0000-00

中文篇章级句间语义关系体系及标注*

张牧宇[1]，秦兵[1]，刘挺[1]

（1.哈尔滨工业大学，黑龙江省 哈尔滨市 150001）

摘要：篇章句间关系（Discourse Relation)是篇章级语义分析的重要内容，本文在英文篇章句间关系研究的基础上分析了中英文间的差异，总结了中文篇章级语义分析的特点，并在此基础上提出面向中文篇章句间关系的层次化语义关系体系，对句间关系类型进行详细描述。为了验证体系的合理性和完备性，我们在互联网新闻语料上进行了标注实践，分析了标注中遇到的难点并给出解决方案，为进一步的中文篇章级语义分析工作奠定基础。

关键词：中文篇章级语义分析；句间关系；语义体系；语料标注；

中图分类号：TP391　　**文献标识码：**A

Chinese Discourse Relation Semantic Taxonomy and Annotation

Muyu Zhang[1], Bing Qin[1], Ting Liu[1]

(1.Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Discourse Relation is an important part of discourse semantic analysis. This paper analyses the differences between Chinese and English, then presents the first Chinese discourse relation taxonomy based on the English discourse relation researches with explanation in details. Aim to check the rationality of the hierarchy, we conducts annotation experiments on Chinese internet news texts and analyses all difficulties happened during the data annotation together with the resolution to lay a foundation for the future discourse semantic analysis.

Key words: Chinese discourse semantic analysis; discourse relation; semantic taxonomy; data annotation;

1 引言

随着词汇语义、句子语义研究的逐渐成熟，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，篇章句间关系研究（Discourse Relation）也开始受到越来越多的关注。该研究以分析文本块间的因果、比较等语义关联为目标，在缺少篇章级资源和方法的情况下，利用篇章句间关系将词汇语义融合成为篇章级语义信息，成为语义分析的重要解决途径之一，对自动文摘（Marcu, 2000）、自动问答（Girju, 2003）、倾向性分析（Somasundaran et al, 2008；Zhou L et al, 2011）以及文本质量评价（Pitler and Nenkova, 2008）、文本连贯性评价（Lin et al, 2011）等许多 NLP 任务起到了很大的帮助。

根据依赖的核心内容不同，传统的语义分析方法通大致可以分为以下三类：

(1) 以词汇语义为核心的分析理论。

词汇链理论（Lexical Cohesion）通过分析词汇语义构建主题词汇链，利用主题词汇的分布和转移情况分析篇章语义（Jane and Graeme, 1991）。该理论的操作性较强，但以词汇

* **收稿日期：2013.3.13**　　**定稿日期：2013.3.29**

基金项目：国家自然科学基金重点项目（61133012）；国家自然科学基金面上项目（61073126）；国家自然科学基金面上项目（61273321）；国家 863 前沿技术研究项目（2012AA011102）

作者简介：张牧宇（1987——），男，博士研究生，主要研究方向为篇章语义分析，文本挖掘；秦兵（1968——），女，教授，主要研究方向为自然语言处理、信息检索、信息抽取等；刘挺（1972——），男，教授，主要研究方向为自然语言处理、文本挖掘、文本检索等。

为主要分析对象，表现力比较有限；中心理论（Grose、Hoshi and Scott,1995）在给定的句子中跟踪焦点变化，通过定义不同的焦点变化方式来描述篇章结构并获知语义信息。该理论主要关注句子间的焦点转换，对篇章信息关注较少，转移关系类型比较单一。

(2) 以语义资源为核心的分析理论。

框架语义学(FrameNet)通过构建“语义框架”将语义知识转化为计算机词典(Fillmore C.J.，1976)，用词义间的关联反映语义。山西大学刘开瑛教授、李茹教授领导的课题组研究构建了中文框架语义知识库（Chinese FrameNet，CFN)。该理论语义信息丰富，便于计算机使用，但专业性极强，构建过程耗时耗力，很难形成规模；脚本理论将文本语义转化为具体的情景知识，通过对典型场景建模将文本语义具体化（Schank et al.,1977)。该理论构思巧妙，但现实场景无法穷尽，限制了实用性。

(3) 以篇章结构为核心的分析理论。

修辞结构理论（Rhetorical Structure Theory，RST）将语义分析转换为文本间的修辞结构分析，用修辞结构树表示文本语义信息（Mann and Thompson，1988)。该理论非常完善，但关系定义偏近于修辞结构，操作性不强；文本意义段划分（Text tiling）通过统计词汇共现和分布情况识别段落主题，并以主题为单位合并文本段，将篇章转换为多个“意义段”的组合（Marti A. Hearst ，1997)。该理论以段落为分析对象，对分析语义结构帮助较大，但分析过程依赖词汇统计，对语义信息使用较少。

第一类方法最为直观，操作性强，但以词汇语义为核心，表现力略有不足；第二类方法表现力强，但构建过程费时费力，场景无法穷尽限制了实用性。第三类方法兼有表现力和实用性，但偏近于结构分析，不利于操作。在这种情况下 PDTB（Penn Discourse Treebank）（Prasad et al.，2008）应运而生，该理论在 PTB 树库基础上，以篇章关联词语为核心，从语义角度出发构建了篇章句间关系树库，语义信息丰富且操作性强。该理论以篇章关联词为核心，从任务定义、体系设计再到标注方法，都是面向英文展开，考虑到中英文关联词使用和语义关系都存在一些差异，将 PDTB 体系直接平移到中文可能会出现偏差。

本文首次提出面向中文的篇章级句间语义关系体系,将基于篇章级句间关系的语义分析方法应用在中文，通过分析中英文的差异指出中文体系的必要性，详细介绍面向中文的语义关系体系并通过语料标注证明了中文体系的一致性和完备性。余下内容组织如下：第二部分介绍篇章句间语义关系相关工作；第三部分分析中英文的差异，说明中文体系的必要性；第四部分介绍本文提出的中文篇章级句间语义关系体系;第五部分介绍中文篇章级句间语义关系语料标注及问题分析；第六部分给出结论。

2 相关工作概述

篇章句间关系理论包括 RST 和 PDTB 两类，其中 RST 理论（Mann and Thompson，1988）的出现是篇章句间关系研究的第一次高潮。Marcu（Marcu，1997）在 RST 框架下提出了一种篇章树结构的分析方法；随后 Soricut（Soricut and Marcu，2003）使用概率模型结合词汇特征和句法特征分析了句子级的篇章结构；DuVerle（DuVerle，2009）使用丰富的词汇、句法和结构化特征，采用 SVM 方法训练两个独立模型识别文本间的修辞关系；VW Feng（VW Feng and Hirst，2012）以 RST 体系为基础，使用了大量的语言学特征构建了文本分析器。

PDTB 体系提出并开放树库之后（Prasad et al.，2008）出现了一系列相关研究。Pitler（Pitler et al.，2008）通过统计关联词识别显示关系（存在关联词的篇章句间关系），取得了非常好的效果；随后 Piter（Pitler et al.，2009）使用句法信息分析篇章关联词的歧义问题，并对隐式关系识别进行了探索（Pitler et al.，2009)；Wellner（Wellner and Pustejovsky，2007）和 Elwell（Elwell and Baldridge，2008）分别使用了机器学习方法，以关联词为核心进行了关系元素识别。

对于不存在关联词的关系（隐式关系）Lin（Lin et al.，2009）使用四类特征进行了识别，证明句法产生规则最为有效；随后 Lin（Lin et al.，2010）提出了基于 PDTB 体系的篇章分析器，实现了篇章关联词识别、元素识别、关系分类等功能；Wang（Wang et al.，2010）采用核函数方法，结合时序信息识别篇章句间关系，取得了当时的最佳效果；Zhou（Zhou et al.，2010）选择关联词插入隐式关系，并利用该关联词识别隐式关系类型；Chiarcos（Chiarcos，2012）利用事件关系指示词帮助识别篇章句间关系。

目前除英文 PDTB 之外，印度语（Prasad et al.，2008b）、土耳其语（Zeyrek and Webber，2008）、阿拉伯语（Alsaif and Markert，2011）也有类似的篇章句间关系研究出现。在中文上，Xue（Xue，2005）针对篇章关联词进行了研究，随后 Zhou（Zhou and Xue, 2012）采用英文 PDTB 体系标注了部分中文语料，但没有提出适合中文特点的语义关系体系。本文提出面向中文的篇章句间语义关系体系，指出中文体系的必要性，并进行了中文语料标注和分析工作。

3 中英文差异分析

PDTB 已经出现并获得了发展，但我们发现中英文间存在一些明显差异，使得英文体系平移到中文时会出现一些偏差，主要包括以下方面。

3．1 英文体系对中文语义覆盖不完整

“英文重结构，中文重语义”，中文中有许多独特的语义关系在英文 PDTB 体系中没有体现，平移到中文时体系覆盖不完整，会出现信息丢失。

例 1：为了报效祖国，他苦练杀敌本领。

例 2：他不但还清了债，还盖了新房子。

例 1 表示目的语义，例 2 表示递进语义。这两类关系在中文里非常常见，但英文 PDTB 体系没有对应的关系类型，对该类句子不进行标注，遗漏了很大一部分信息。本文通过定义“目的关系”、“递进关系”覆盖以上两类语义信息。

3．2 英文体系对某些关系分类不清

“英文多从句，中文多分句”，中文篇章经常出现多个语义地位相似的连续分句，描述问题的若干方面，语义独立且地位并列，英文 PDTB 体系没有覆盖并列语义的关系类型。

例 3：一方面通过限购和限贷减少市场上的投机性需求;另一方面，增加保障性住房和普通商品房的供应。

例 3 中两个分句说明并列的两方面举措，这种用法在中文非常常见。英文 PDTB 体系缺少对应的关系类型，通常会识别为“扩展关系”等，在后续使用中（例如问答）常常会将后半部作为附属丢弃，导致信息丢失，实际上二者并没有内容上的从属关系。本文通过定义“并列关系”获得元素间的语义并列。

3．3 英文时态关系平移困难

英文 PDTB 体系中包含面向时态信息定义的关系类型，由于中文缺少明显的时态和从句信息，相关类型平移到中文时会导致识别困难，并影响信息间的区分方式。

例 4：If they had this much trouble with Chicago, they are going to have an awful time with the rest.

例 4 属于英文 PDTB 体系中“条件关系”下属的“过去事实”，定义中使用了动词的时态信息。类似信息在中文里识别困难，而且依靠时态区别各类关系会隐藏信息间的其他区别。本文提出的中文体系删除了时态相关关系类型，并采用其他区分标准重构条件关系。

3．4 英文元素定义在中文上的使用困难

根据传统的任务定义（Prasad et al.，2008），篇章句间关系包括两个元素，在偏正复句中两者功能不同（王力，1984），需要在识别过程中加以区分。英文 PDTB 体系根据关联词

的修饰关系区分元素：关联词修饰的元素称为 Arg2，另一个元素称为 Arg1。

例 5：如果做完了功课（Arg2），你可以出去玩一会（Arg1）。

例 6：做完了功课（Arg1），你就可以出去玩一会（Arg2）。

这种定义以关联词为核心，配合英文对关联词的严格限制可以满足英文的处理要求。但该定义平移到中文时可能出现一些偏差，主要包括以下方面：

(1) 平行关联词

英文复句通常只允许一个关联词出现，而中文复句允许多个关联词语同时出现，两个元素均有关联词修饰，按照英文定义无法区分 Arg1 和 Arg2。

例 7：因为河里结了冰，所以轮船开不出去了。

(2) 关联词换位

英文关联词使用比较严格，以条件关系为例：关联词通常只出现在条件分句中（例 5），因此英文 PDTB 条件关系只能覆盖 Arg1 描述结果的实例，忽略 Arg1 描述条件的情况（例 6）。中文关联词经常在句子中换位出现，平移英文定义会丢失例 6 中的语义关系。

(3) 语义歧义

同一个语义内容可以使用不同的关联词（例 8,9）。按照英文 PDTB 的定义方式，同一个句子在语义内容不变的情况下可以对应两种关系，这会导致更多的语义歧义。

例 8：河里结了冰，所以轮船开不出去了。（原因关系）

例 9：因为河里结了冰，轮船开不出去了。（结果关系）

本文提出基于物理顺序的元素定义方法：将前置的元素设为 Arg1，后置的为 Arg2。这种定义方法识别简便，适用任何语言变化，在平行关联词、关联词换位和隐式关系处理上优势明显；此外，在这种定义下例 8,9 将具有相同的关系类型，更加符合语义逻辑。

4 中文篇章句间语义关系体系

“英文是法制的，中文是人治的”（王力，1984）。中英文的差异较大：从语义表现上讲，中文缺少时态、从句等标识，句子间往往通过语义关联，关系更为丰富；从表达方式上讲，中文关联词的使用更为随意，行文非常灵活；从组织方式上讲，中文没有明显的修饰结构，句子分割灵活，元素构成更加复杂，因此直接使用英文体系处理中文存在一定偏差，有必要提出面向中文的篇章语义关系体系。

4．1 概念介绍

类似于传统篇章句间关系任务（Prasad et al.，2008），中文篇章句间语义关系体系包含三个层次内容：篇章句间关系、关系元素、语义关系体系。

篇章句间关系：两个文本块间存在的语义关联，通常由关联词进行标识。根据关联词语不同可以分为以下三类：

(1) 显示关系：存在明显的关联词语作标识的篇章句间关系。

(2) 隐式关系：缺少关联词语，需要进行推理识别的篇章句间关系。

(3) 替换关系：缺少关联词语，但具有功能类似的短语或其他成分的。

关系元素：构成篇章句间关系的文本单元。根据关系类型不同可以分为以下三类：

(1) 显示关系元素：显示关系元素以语义独立为主、关联词为辅，允许嵌套情况，组成单位包括简单句、短语、复句、指代表达式四类。

(2) 隐式关系元素：英文隐式关系只能位于段内的相邻句之间，考虑到中文跨句关系较多，中文体系取消了这个限制，隐式关系元素定义同显示关系。

(3) 替换关系元素：本文按照显示关系处理替换关系，元素定义同显示关系。

语义关系体系：篇章句间关系所有可能类型组成的层级体系，下面进行详细介绍。

4．2 篇章句间语义关系体系

中文篇章句间语义关系体系具有多层级结构：每个一级类目都包括若干二级类目，部分二级类目可以细化为三级类目。这种层级结构符合多标注一致性规则，保证定义的精确性，并且支持一些非常实用的推理。本文用 Arg1 标识位置前置的元素，Arg2 标识位置后置的元素，用下划线标识关联词语。

4.2.1 时序关系

两个元素是通过时序联系起来，称之为“时序关系”。

- 同步关系

两个元素存在一定时间重合，称之为“同步关系”。

例 10：当[他打篮球]$_{arg1}$的时候，[天突然下起了雨]$_{arg2}$。

- 异步关系

两个元素存在时间上的先后顺序，称之为“异步关系”。

  - 先序关系

两个元素中 Arg1 在 Arg2 之前发生，称之为“先序关系”。

例 11：在[上课铃打响]$_{arg1}$之后，[他才跑回了教室]$_{arg2}$。

  - 后序关系

两个元素中 Arg1 在 Arg2 之后发生，称之为“后序关系”。

例 12：在[上课铃打响]$_{arg1}$之前，[他飞快的跑回了教室]$_{arg2}$。

4.2.2 条件关系

两个元素中一个提出某种条件或场景，另一个说明产生的结果，称之为“条件关系”。

- 直接条件关系

两个元素分别描述条件和结果，两者存在直接的条件影响，称之为“直接条件关系”。

  - 充分条件

条件元素为真时，结果元素一定为真，称之为“充分条件”。

例 13：只要[你通过了初试]$_{arg1}$，就[能参加面试了]$_{arg2}$。

  - 必要条件

结果元素为真时，条件元素一定为真，称之为“必要条件”。

例 14：只有[赢下这场比赛]$_{arg1}$，[他们]才能进入下一轮]$_{arg2}$。

  - 无条件

无论条件元素是否为真，结果元素都将发生，称之为“无条件”。

例 15：无论[明天是否晴天]$_{arg1}$，[我们]都将上路]$_{arg2}$。

- 形式条件关系

两个元素中一个提出某种场景，另一个以此为背景展开内容，不存在直接的条件影响，而是形式上的铺垫，称之为“形式条件关系”。

例 16：如果[谁要鉴赏我国的园林]$_{arg1}$，[苏州园林就不该错过]$_{arg2}$。

4.2.3 比较关系

两个元素描述的内容存在某些差异，并且突出这种差异时，称之为“比较关系”。

- 直接对比关系

两个元素存在某些相似的属性，突出两者在该属性上的差异，称之为“直接对比关系”。

  - 同向对比

两个元素在公用属性上具有同趋势的取值，称之为“同向对比”。

例 17：2009 年[公司盈利提升了 69%]$_{arg1}$。但是[同年互联网方面的开销增加 85%]$_{arg2}$。

  - 反向对比

两个元素在公共属性上具有不同趋势的取值，存在一定语义转折，称之为“反向对比”。

例 18：[公司的销售额提升了]$_{arg1}$。但是[利润却没有增加]$_{arg2}$。

- 间接对比关系（转折关系）

两个元素没有明显的属性对比，而是语气转折或话题转换，称之为“间接对比关系”。

例 28：[中国的奢侈品消费增长如此之快]$_{arg1}$。但是[消费者的眼光也越来越挑剔了]$_{arg2}$。

4.2.4 因果关系

两个元素间存在事实性的因果影响，称之为“因果关系”。

- 直接因果关系（说明因果）

两个元素中一个说明原因，另一个说明该原因导致的结果，两者均为客观事实，称之为“直接因果关系”。

  - 原因关系

两个元素中 Arg1 说明原因，Arg2 说明该原因导致的结果，称之为“原因关系”。

例 19：[他昨天淋了雨]$_{arg1}$。 所以[今天有点发烧]$_{arg2}$。

  - 结果关系

两个元素中 Arg2 说明原因，Arg1 说明该原因导致的结果，称之为“结果关系”。

例 20：[他不能来开会了]$_{arg1}$，因为[他生病了]$_{arg2}$。

- 间接因果关系（推论因果）

两个元素中一个提出某些依据，另一个由此得出主观结论，称之为“间接因果关系”。

例 21：[警察认定他撒了谎]$_{arg1}$。[他们在他的住所找到了失主的钱包]$_{arg2}$。

- 目的关系

两个元素中一个提出某种目标，另一个指出该目标所需的行动，称之为“目的关系”。

例 22：[为了让老师相信他]$_{arg1}$，[他写了八百多字的保证书]$_{arg2}$。

4.2.5 扩展关系

两个元素间存在内容上的扩展，推动行文向前，称之为“扩展关系”。

- 细化关系

两个元素中一个是对另一个的展开和细化描述，称之为“细化关系”。

  - 解释说明

两个元素中 Arg2 是对 Arg1 的进一步解释，称之为“解释说明”。

例 23：[这是一个很古老的故事]$_{arg1}$。[它发生在 500 年前那个混乱的年代]$_{arg2}$。

  - 实例关系

两个元素中一个覆盖了某个集合，另一个描述该集合的特定实例，称之为“实例关系”。

例 24：[今年他获得了很多大奖]$_{arg1}$。其中一个[是季度之星]$_{arg2}$。

  - 例外关系

两个元素中一个提出某个理论，另一个指出例外情况，称之为“例外关系”。

例 25：[中国男足一直无法突破世界杯预选赛]$_{arg1}$，除了[02 年韩日世界杯]$_{arg2}$。

- 泛化关系

两个元素中 Arg2 是对 Arg1 的总结和泛化时，称之为“泛化关系”。

例 26：[这个蒸锅可以蒸玉米、番薯、热饭啥的，还可以煲汤]$_{arg1}$，总之[用处蛮大滴]$_{arg2}$。

- 递进关系

两个元素间是层进关系，Arg2 比 Arg1 更进一层，称之为“递进关系”。

例 27：[他不但表达能力强]$_{arg1}$，而且[动手能力也相当不错]$_{arg2}$。

4.2.6 并列关系

两个元素的内容和语义相互并列，处于同等地位，称之为“并列关系”。

- 平行关系

两个元素分别陈述几个问题，或同一问题的几个方面，称之为“平行关系”。

例 28：[近来开发商资金越来越紧张]$_{arg1}$，同时[政府的调控也在加强]$_{arg2}$。

- 选择关系

两个元素描述一种选择情况，称之为“选择关系”。

- 相容选择

两个元素均可供选择并且都鼓励尝试，称之为“相容选择”。

例 29：建议你[平时多多读书]$_{arg1}$，或者[多和其他人交流]$_{arg2}$。

- 互斥选择

两个元素只能二中选一，不能同时出现，称之为“互斥选择”。

例 30：要么[骄傲的死去]$_{arg1}$，要么[耻辱的活着]$_{arg2}$。

5 语料标注实践与分析

为了检验中文篇章句间语义关系体系的合理性和完备性，本文首次进行了中文体系上的语料标注工作。我们认为：如果多名标注人员的一致性较高，说明语言实例到中文体系的映射比较一致，从而证明中文语义关系体系一致性较好；如果多名标注者完全不同的实例较少，说明绝大多数关系实例都存在对应的合适类型，从而证明中文语义关系体系的完备性。

5．1 标注内容

显示关系标注：要求识别所有存在关联词标识的篇章句间关系并确定类型，这里的关联词除常用词外，还包括“面对着”、“一上来”、“其中一个”等短语活用现象。

隐式关系标注：要求识别所有不存在关联词标识的篇章句间关系，包括相邻和跨句两种情况，在识别之后需要标注人员选择合适的关联词插入句子标识该关系。

元素标注：要求识别构成篇章句间关系的文本块，包括短语、分句、分句组合、复句、句群等，并且允许嵌套。

5．2 语料来源及标注设置

本次标注使用互联网新闻语料，来源包括新浪和搜狐两个门户网站。在网站新闻频道下属的各个顶级类别下各选择一篇质量较好的语料，并分为两组用于培训和测试。数据标注由三名标注者执行，均有自然语言研究背景，但均无篇章句间关系研究经历。

为了证明中文体系的一致性，本文把语义标注看成分类问题，采用标准 Kappa 值（公式 1）计算多名标注人员的标注一致性，并统计了多名标注者结果完全相同的实例数目。

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

注：P（A）表示多名标注者评定一致的百分比，P（E）表示理论上评定一致的百分比

为了证明中文体系的完备性，本文统计了多名标注人员结果完全不同的实例数目。如果这种强歧义的关系实例较多，说明现有体系中缺少合适的关系类型；反之则说明现有体系能覆盖绝大多数关系实例。

5．3 标注分析

本文采用严格匹配的方法计算三名标注者（A1、A2、A3）的标注一致性 kappa 值：只有多个标签全部细化到底层类别，并且结果一致才认为结果相同，详细结果如表 1 所示。

分析下表可知：A2 与 A3 标注一致（A2=A3）的 Kappa 值最高，达到了 74.10%； A1 与 A3 标注一致（A1=A3）的 Kappa 值为 66.70%；而三名标注者全部一致的 Kappa 值达到 59.90%，整体 Kappa 值较高。根据之前分析，较高的标注一致性说明中文篇章句间语义关系体系一致性较好。此外三名标注者结果完全不同的实例极少，说明绝大部分实例都可以划入合适的类别，从而证明中文体系的完备性。

另外值得注意的是三个标注者彼此间标注一致性差异较大：A2 与 A3 标注一致（A2=A3）的 Kappa 值达到了 74.10%，而 A1 与 A2 标注一致（A1=A2）的 kappa 值只有 59.30%；整

体而言，标注人员 A1 与 A2、A3 的一致性低于 A2 与 A3 的一致性。这从一个侧面说明了语义问题的复杂性，标注结果一定程度上依赖于标注人员。语义问题本身具有歧义，再加上中文的变化较多，很难应用统一的标准判别，为自动处理增加了难度。

表 1 标注结果分析

Tab.2 Analysis of the data annotation

类别	传媒	娱乐	IT	汽车	教育	Kappa
关系数目	9	8	4	16	18	
A1=A2	4	4	3	12	9	59.30%
A1=A3	4	5	4	13	10	66.70%
A2=A3	6	6	3	13	12	74.10%
全部一致	4	5	3	13	12	59.90%
全部不同	0	0	1	1	2	

注： 三名标注人员 A1、A2、A3 分别标注 55 个篇章句间关系实例；之后先将三组标注结果两两组合，计算组合内的标注一致性；然后计算三名标注者的整体标注一致性，共得四组 Kappa 值

5．4 标注问题分析

语义问题本身具有一定的歧义性，再加上中文组织方式灵活，使得中文篇章句间关系分析更加困难。下面分析标注过程中遇到的问题，并给出解决方案。

5.4.1 信息缺失

中文短句较多，很多分句的句法成分并不完整（例 31 加粗部分），但具有相对完整的语义（例如省略、指代、特定表达），是否允许它们作为元素参与构成关系值得我们思考。

例 31：谷歌向开发人员提供这些工具还有什么意义呢？谁又愿意去开发那些需要 ICS 功能的应用呢，**更何况是 Android 5.0**？

目前考虑解决方案如下：

- 允许作为元素出现
- 首先判定句子语义是否完整，忽略内容缺失的句子

5.4.2 句法与语义关系区别

中文允许分句作为句法成分，例 32 中“指出”以分句作为宾语，对行文展开有一定推动作用，类似的句法关系是否属于篇章句间关系的处理范围值得我们思考。

例 32：国外媒体今日撰文**指出**，谷歌每年推出一款重大升级的做法似乎并不恰当。

目前考虑解决方案如下：

- 添加相应的句法关系类别
- 只关注语义问题，句法关系交由信息抽取任务处理

5.4.3 语义歧义

语义问题具有一定歧义性，加之中文语义表达方式丰富，使得很多关系类别的判定依赖于标注人员。例 33 加黑部分介于“递进关系”与“解释说明”之间，如何标注类似的歧义关系值得我们思考。

例 33：Android 4.0 操作系统(代号“冰淇淋三明治”)自发布以来，仅占整个 Android 市场份额的 1%。**目前仅有一少部分手机和平板电脑采用 Android 4.0。**

目前考虑解决方案如下：

- 对于某些歧义性较强关系实例，允许属于多个类别

6 结论

本文首次提出了面向中文的篇章句间语义关系体系，对中文篇章句间关系研究进行了初步尝试，基于中文体系进行了新闻语料标注，分析了标注过程中遇到的问题并给出解决办法，对中文篇章句间关系研究进行了初步尝试。

本文提出的篇章句间语义关系体系符合中文语言学相关理论，通过语料标注和多名标注人员的标注一致性分析验证了中文篇章句间语义关系体系的合理性，同时反映语义问题自身的歧义性。通过对比中英文的差异指出英文体系平移时可能遇到的问题，指出中文体系的必要性，并提出针对性的调整方案。

在此基础上，我们计划开展以下工作内容：中文篇章句间关系语料标注及语料库构建；中文篇章句间关系类型的自动识别方法；中文篇章句间关系元素的定义域自动识别方法；文本多级语义关系嵌套问题的分析与解决；文本语义分析器的设计与实现。

7 致谢

本文获得国家自然科学基金重点项目（61133012）、国家自然科学基金面上项目（61073126）、国家自然科学基金面上项目（61273321）以及国家 863 前沿技术研究项目（2012AA011102）资助，特此感谢。

参考文献

[1] D. Marcu. 2000. The theory and practice of discourse parsing and summarization. MIT Press.
[2] R. Girju. 2003. Automatic detection of causal relations for questions answering. In Proc. of the ACL 2003 Workshop on Multilingual Summarisation and Question Answering, pages 76-83.
[3] S. Somasundaran, J. Wiebe, and J. Ruppenhofer. 2008. Discourse-level opinion interpretation. In Proc. Of Coling 2008.
[4] E. Pitler and A. Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In Proc. of EMNLP 2008, pages 186-195.
[5] E. Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Singapore, 2009.
[6] E. Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In 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, 2009.
[7] Ziheng Lin, Hwee Tou NG, and Min-Yen Kan. 2011. Automatically Evaluating Text Coherence Using Discourse Relations. s. In Proceedings of ACL-HLT, pages 997-1006.
[8] Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 2009.
[9] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A PDTB-styled end-to-end discourse parser. Technical Report TRB8/10, School of Computing, National University of Singapore, August.
[10] Zhou, L., Li, B., Gao, W., Wei, Z., Wong, K.F. 2011. Unsupervised Discovery of Discourse Relations for Eliminating Intra-sentence Polarity Ambiguities. In Proceedings of EMNLP 2011 (Oral presentation), Edinburgh, Scotland, July 27-31.
[11] Morris J, Hirst G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1):21-48.
[12] Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A Framework for Modelling the Local coherence of Discourse. Computational Linguistics, 21/2: 203 - 25.
[13] William C. Mann and Sandra A. Thompson. Relational propositions in discourse (A). Information Sciences Institute of University of Southern California, Los Angeles, 1893.
[14] Marti A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1):33-64.
[15] Fillmore, Charles J. 1976. Frame semantics and the nature of language. In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, Volume 280: 20 - 32.

[16] Schank, R C, A belson, R. Scripts, Plans, Goals, and Understanding [M]. Hillsdale, N J: Earlbaum Assoc, 1977.
[17] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008a. The Penn discourse treebank 2.0. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).
[18] David duVerle and Helmut Prendinger. A novel discourse parser based on Support Vector Machine classification. In 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, 2009.
[19] VW Feng and Hirst. Text-level Discourse Parsing with Rich Linguistic Features. In Proc. Of ACL 2012.
[20] BonnieWebber and Aravind Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. In Manfred Stede, Leo Wanner, and Eduard Hovy, editors, Discourse Relations and Discourse Markers: Proceedings of the Conference, pages 86-92. Association for Computational Linguistics.
[21] Ben Wellner and James Pustejovsky. Automatically identifying the arguments of discourse connectives. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, 2007.
[22] Robert Elwell and Jason Baldridge. Discourse connective argument identification with connective specific rankers. In Proceedings of the IEEE International Conference on Semantic Computing, Washington, DC, USA, 2008.
[23] Z. Zhou, Y. Xu, Z. Niu, M. Lan, . Su, and Tan. C. 2010. Predicting discourse connectives for implicit discourse relation recognition. In Proc. of Coling 2010, pages 1507-1514.
[24] C. Chiarcos. Towards the Unsupervised Acquisition of Discourse Relations. In Proc. Of ACL 2012.
[25] WenTing Wang, Jian Su, and Chew Lim Tan. 2010. Kernel based discourse relation recognition with temporal ordering information. In Proceedings of the 48th Annual Meeting of the Association for Computatio
[26] Daniel Marcu. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. PhD thesis, University of Toronto, 1997.
[27] Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, Canada, 2003.
[28] Rashmi Prasad, Samar Husain, Dipti Sharma, and Aravind Joshi. 2008b. Towards an annotated corpus of discourse relations in Hindi. In Proceedings of the Third International Joint Conference on Natural Language Processing, Hyderabad, India.
[29] Deniz Zeyrek and Bonnie Webber. 2008. A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. In Proceedings of IJCNLP-2008. Hyderabad, India.
[30] Amal Al-Saif and Katja Markert. 2011. Modelling discourse relations for Arabic. In Proceedings, Empirical Methods in Natural Language Processing, pages 736-747.
[31] Xue, Nianwen. 2005. Annotating discourse connectives in the Chinese Treebank. In Proc. of The ACL Workshop in Frontiers in Annotation II: Pie in the Sky. Ann Arbor, Michigan: ACL,.
[32] Yuping Zhou, Nianwen Xue. PDTB-style Discourse Annotation of Chinese Text. In Proc. Of ACL 2012.
[33] 王力. 《王力文集》. 山东. 山东教育出版社. 35-36 页.

第一作者
姓名：张牧宇　邮编：150001
电话：+86 13796055168　电子邮箱：myzhang@ir.hit.edu.cn
地址：哈尔滨市南岗区教化街 29 号六楼 社会计算与信息检索研究中心
第二作者
姓名：秦兵　　邮编：150001
电话：0468 86413683　　电子邮箱：qinb@ir.hit.edu.cn
地址：哈尔滨市南岗区教化街 29 号六楼 社会计算与信息检索研究中心
第三作者：
姓名：刘挺　　邮编：150001
电话：0468 86413683　　电子邮箱：tliu72@vip.126.com
地址：哈尔滨市南岗区教化街 29 号六楼 社会计算与信息检索研究中心