

汉语篇章小句关联结构的表示与识别研究

冯文贺^{1,†} 陈伊琳¹ 任亚峰² 任函¹

1. 广东外语外贸大学语言工程与计算实验室, 广州 510006; 2. 广东外语外贸大学外语研究与语言服务协同创新中心, 广州 510006;

† 通讯作者, E-mail: wenhefeng@gmail.com

摘要 将篇章结构表示为小句关联结构,用以刻画篇章中小句之间的直接语义关联,比之层次化篇章结构模式,它可以有效刻画非连续和跨层级的小句之间的语义关联。首先,提出篇章小句关联结构的形式表示、判断准则及形式限制,进行人工标注实验。其次,研究汉语篇章小句关联结构的自动识别。在自建汉语篇章小句关联结构语料库上,基于分类模型,设计连接词、词汇等特征,最佳识别准确率达 92.70%。实验发现:语料整体取样较独立取样取得的去环效果好;词汇、小句距离及句域等特征对识别贡献较大;远距离和跨大句是小句关联识别的难点,但相邻小句和同一大句内的小句对不相关识别则尤其难。

关键词 小句关联结构; 篇章结构; 篇章依存结构; 修辞结构

中图分类号 TN914

Representation and Recognition of Clauses Relevance Structure in Chinese Text

FENG Wenhe^{1,†}, CHEN Yilin¹, REN Yafeng², REN Han¹

1. Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510006; 2. Collaborative Innovation Center for Language Research and Service, Guangzhou 510006;

†Corresponding Author, E-mail: wenhefeng@gmail.com

Abstract In this paper, the discourse structure is represented as the clause relevance structure, which is used to describe the direct semantic association between the clauses in a text. Compared with the hierarchical discourse structure pattern, the semantic association between discontinuous and cross-level clauses can be effectively described. Firstly, the scheme of clause relevance structure and its judgment criteria, formal constraints and manual annotation experiments are proposed. Secondly, the automatic recognition of Chinese clause relevance structure is studied. On the corpus of Chinese discourse clause relevance structure we built, the best recognition accuracy is 92.70% based on the classification model, with the design of connectives, vocabulary and other features. The experimental results show that the ring-removing effect obtained by the overall sampling of corpus is better than that of independent sampling, and the features of vocabulary, clause distance and clause domain contribute greatly to the recognition. Long distance and cross sentences of clause pair are the difficulties of clause relevance recognition, but adjacent clauses and clauses in the same sentence are especially difficult to recognize as uncorrelated clauses.

Key words clause relevance structure; discourse structure; discourse dependency structure; rhetorical structure

篇章结构分析主要分析篇章单位间的结构关系,在自动摘要、自动问答及机器翻译等研究中有重要应用。比较著名的篇章结构表示模式主要有修辞结构^[1,2]和宾州语篇树库模式^[3],前者将篇章表示为一颗层次化结构树,后者将篇章表示为一个独立的篇章连接词的论元结构,其中论元大致相当于篇章单位。本质上,二者都是一种层次化结构,不同在于前者是完整篇章结构树,后者据连接词的论元辖域可推导出一定篇章结构树,但并不构造完整结构树。Li 等^[4]所提篇章连接依存树模式融合了修辞结构的层次化模式与宾

国家社科基金“汉语篇章结构的特征-依存描写机制及资源建设研究”(17BYY036)资助

收稿日期: 0000-00-00; 修回日期: 0000-00-00; 网络出版时间:

网络出版地址:

州语篇树库的连接词模式，在获得完整的篇章结构树同时兼表示连接词论元结构，在本质上仍然是一种层次化结构。层次化结构的问题在于不刻画最小篇章单位间的直接语义关联，特别是不能刻画非相邻和跨层级的最小篇章单位间的语义关联，而这种关联在篇章中客观而普遍存在。

篇章依存结构^[5,6,7]为解决以上问题提供了形式表示的可能性。类似传统句法依存结构^[8]要确定词与词之间的中心与非中心关系，篇章依存结构的核心在于确定最小篇章单位之间的中心与非中心关联。然而，如何确定最小篇章单位间有无关联，何为中心，这些基础理论问题至今缺少内涵性研究。而相应的篇章依存语料库工作或由修辞结构语料库简单转化而来^[5]，或关系体系等略调面向特定领域^[6]，且均非基于汉语。

本文将篇章结构表示为小句关联结构，小句^[9]即最小篇章单位。比之篇章层次结构，它可以有效刻画非相邻和跨层级的最小篇章单位之间的语义关联。小句关联与现有篇章依存结构相关但不同，现有篇章依存结构一般确定小句关联的同时确定中心，而小句关联结构仅关注小句之间有无关联，相比更纯粹。本文的小句关联结构上也可构拟中心，但并不在构建小句关联的同时确定中心，而是在确定所有小句关联结构之后，才基于小句关联结构整体拟定中心。如果将进一步区分了中心的小句关联结构也称为“篇章依存结构”，那么在这种篇章依存结构分析中，小句关联是第一性的，关联中心是第二性的，小句关联分析更具基础性。

本文提出篇章小句关联结构表示机制，阐明小句关联的判断准则、形式限制等，进行语料标注实验（第1节）。在自建的汉语篇章小句关联结构标注语料上，进行探索性小句关联识别研究，重点考察计算模型、语料取样、语言特征等对小句关联识别的影响（第2节）。

1 篇章小句关联结构

1.1 形式表示

小句（clause）是最小篇章单位，两个有关系的小句构成一个小句关联对，即为“小句关联”。图1为例1的篇章小句关联结构图式，数字“1、2”等代表小句序列，小句关联由连线表示。小句关联结构整体上是一个无向无环图。

例1.¹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，²因此大量出现的是以前不曾遇到过的新情况、新问题。³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，⁴而是借鉴发达国家和深圳等特区的经验教训，⁵聘请国内外有关专家学者，⁶积极、及时地制定和推出法规性文件，⁷使这些经济活动一出现就被纳入法制轨道。⁸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，⁹正因为一开始就比较规范，¹⁰运转至今，¹¹成交药品一亿多元，¹¹没有发现一例回扣。

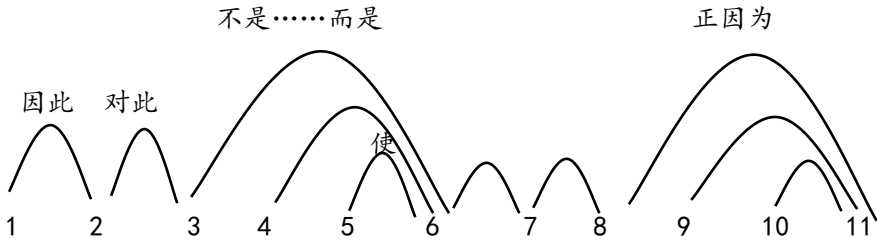


图1 篇章小句关联结构[例（1）]
Fig.1 Discourse clauses relevance structure [Example (1)]

相比修辞结构等层次化结构，小句关联突破篇章单位的线性顺序和层级限制，准确刻画了基本篇章单位间直接关系，如 3-6、4-6、8-11、9-11 突破篇章单位线性顺序限制而直接关联，2-3、7-8 突破篇章单位不在同一大句（sentence，由句号等句末点号标记，例中用“一、二”等标记）的层级限制而直接关联。

基于小句关联及其度量，一般可以进一步推关联结构的中心及全文主旨。如例1中，小句6因为与多个小句（3、4、5、7）直接关联及与一些小句相对间接关联，可推出其为相关关联对（3-6、4-6、5-6、6-7）的中心，进一步又可推导出其为全段中心，这种关联中心，相比修辞结构的关系中心，更具全局性，也更具篇章本质。基于小句关联结构，可以更准确刻画篇章语义关系，如连接词可以突破线性位置的连续性和

单位层级限制，跨篇章单位，直接反映小句关联间关系，如“不是……而是”“正因为”分别反映非连续篇章单位 3-6、8-11 间关系，“对此”反映跨大句的小句 2-3 间关系。相比而言，连接词反映的小句关联关系，比在层次化结构中反映的语义关系更具体、准确。小句关联与基于其上的篇章中心与语义，即为我们进一步构拟的“篇章依存结构”（另文全面探讨）。在篇章依存结构分析中，小句关联分析更具基础性。由于独特性，小句关联分析在自动摘要、自动问答及机器翻译等研究中有独特的应用价值。

1.2 判断准则

直观上，关联小句可构成仅含两个小句的最小复句，这是小句关联的直观判断机制。两个关联小句如果要成为自然语篇，还要删除或添加无关或有关衔接成分。例 2 给出例 1 的小句关联（部分），删除了当前关联句对中无关的连接词（删除号标出），又添加了一些相关的主语成分等（括号标出）。

例 2. 1-2: ¹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，²因此大量出现的是以前不曾遇到过的新情况、新问题。

2-3: ²因此（浦东开发开放）大量出现的是以前不曾遇到过的新情况、新问题。³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，

3-6: ³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，⁶而是……积极、及时地制定和推出法规性文件，

4-6: ⁴而是（浦东）借鉴发达国家和深圳等特区的经验教训，⁶积极、及时地制定和推出法规性文件，

5-6: ⁵（浦东）聘请国内外有关专家学者，⁶积极、及时地制定和推出法规性文件，

6-7: ⁶（浦东）积极、及时地制定和推出法规性文件，⁷使这些经济活动一出现就被纳入法制轨道。

……

在理性机制上，小句关联判断的依据在于：第一，语义连贯。对小句间语义连贯的判断，在根本上，可归结为小句间词汇、短语等构成要素的语义关联。例 2 用下划虚线标出小句关联间某些有语义关联的词汇、短语。如 1-2 中，小句 1 的“跨世纪工程”与小句 2 的“新情况、新问题”有某种关联。一个小句与不同小句有关联，往往是这个小句的不同部分与不同小句关联，如小句 3，在 2-3 中，主要是小句 3 的“采取……做法”与小句 2 的“情况、问题”关联（因果性关联），在 3-6 中，则主要是小句 3 的“干一段时间，等积累了经验以后再……”与小句 6 的“积极、及时……”构成某种关联（转折性关联）。

第二，形式衔接。两个小句关联，通常要求依附于小句的指代、连接词等衔接成分，可从关联小句上得到解读，或关联小句本身语义自足，无指代等需要解读。试建立例 1 中可能的两个小句关联 2-3、2-6：

例 3. 2-3: ²因此（浦东开发开放）大量出现的是以前不曾遇到过的新情况、新问题。³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，（✓）

2-6: ²因此（浦东开发开放）大量出现的是以前不曾遇到过的新情况、新问题。⁶（浦东）积极、及时地制定和推出法规性文件，（×）

对比发现，两者都语义连贯，甚至 2-6 更自然。但最终选择 2-3，而非 2-6，这是因为，2-6 中，小句 6 的主语（“浦东”）为零指代，需要得到解读，而小句 2 的主语为“浦东开发开放”，不能从小句 2 得到较好解读；而 2-3 中，小句 3 本身语义自足，无指代需要解读。相比之下，2-3 的衔接性好，由此选择 2-3 关联，而不选择 2-6 关联。

那么，小句 6 的主语（“浦东”）从何处解读？往前看，只能从小句 3 的主语（“浦东”）得到解读，由此，小句 6 与小句 3 构成 3-6 关联。又则，“不是……而是”作为连接词，需要从语义上得到解读。表面上“不是”在小句 3，“而是”在小句 4，然而 3-4 关联的语义内容显然不符合“不是……而是”的语义表示（表“反转”）。对比之下，3-6 关联的语义内容非常符合“不是……而是”的语义，小句 3 “干一段时间，等积累了经验以后再……”与小句 6 “积极、及时……”，正好语义相反。由此，3-6 关联可使其间的指代、连接词等均得以解读。

1.3 形式限制及语义解释

篇章中所有小句联通，但小句的联通不可有环。这是篇章整体小句关联结构的形式限制。对于 n 个小句构成的语篇，小句关联数目(X)为 $n-1$ 。当 $X < n-1$ ，小句不能完全联通；当 $X > n-1$ ，小句联通产生环。

小句均联通是篇章整体性的形式要求，小句关联无环是语义表示的最简性要求使然。难点在小句关联无环的限制。如例 1 中，小句关联 4-5、4-6、5-6 单个从语义连贯和形式衔接看，均可接受：

例 4. 4-6: ⁴而是（浦东）借鉴发达国家和深圳等特区的经验教训, ⁶积极、及时地制定和推出法规性文件,

5-6: ⁵（浦东）聘请国内外有关专家学者, ⁶积极、及时地制定和推出法规性文件,

4-5: ⁴而是（浦东）借鉴发达国家和深圳等特区的经验教训, ⁵（浦东）聘请国内外有关专家学者,

但均建立其间关联,从整体上则形成一个有环的结构。这个时候一般从语义的直接关联和间接关联上进行辨析,以消除其中的关联环。这里认定,4-6、5-6 为直接关联,4-5 为间接关联。区分依据在于,4-6、5-6 有内在的逻辑联系,即 4、5 均是 6 的某种实现条件;而 4-5 间并无内在的逻辑联系,只是相对于 6 而言,有着同样的地位,即形成通常所谓“并列”。故此,4-6、5-6 为直接关联,4-5 为间接关联,间接关联可以依据直接关联得到一定推导。同一语篇中的所有小句皆有关联,然其关联或直接或间接,小句关联结构仅关注直接关联。这正如,世间万物皆有关联,这是绝对的,而无关联只是相对而言。

1.4 标注一致性

基于以上理论研究,进行语料库标注一致性实验。对两名中文本科生进行训练,主要结合例证说明判断准则及形式限制等,又进行数篇语料的标注校正指导。采用汉语宾州树库^[10]的新闻语料,开发相应标注平台与评估平台上,进行双盲标注与评估。在 20 篇新闻语料上,小句切分^①的 kappa 值为 94.6%,小句关联的 kappa 值为 90.1%。标注实验表明,篇章小句关联结构分析可行性、科学性较强。

2 小句关联自动识别

将小句关联识别当成一个判断任意两个小句有无关联的分类问题。在自建的汉语小句关联语料库上,利用常见分类模型,采用相关特征训练分类器,考察分类器、语料取样、分类特征等对于自动识别的影响。

2.1 分类特征

所采用分类模型均为常见模型(见 2.3),不再介绍,重点说明分类特征(Feature)的设计及实现。

(一) 连接词

F1: 两个小句包含的连接词及其位置。F2: 两个小句包含的连接词是否为一对关联词。

连接词连接两个关联小句,不同连接词结构位置不一样,对关联小句的分布方向也有不同预测。如例 1 中,小句 2 上的连接词“因此”,预测小句 2(结果)的关联小句(原因)在前,而不在后;而小句 8 上的连接词“(正)因为”预测小句 8(原因)的关联小句(结果)在后,而不在前。关联词由不同词对构成,如“尽管……但”,往往预测各自所在的小句相关联。对于连接词的识别及是否关联词的识别,主要利用汉语篇章结构语料库(CDTB1.0)^[4]的连接词信息。

(二) 词汇

F3: 两个小句包含的相同词及词性。F4: 两个小句包含的相同词个数。F5: 两个小句包含的同义词个数。F6: 两个小句包含的同类词个数。F7: 两个小句包含的相同词、同义词、同类词个数。

相关联小句通常描述相关事物或动作。两个小句中出现相同词,且为实词(名、动、形、副),则两个小句的关联程度大。两个小句中出现相同词、同义词、同类词越多,两个小句的关联程度越大。其中,同义词、同类词确定根据哈工大《同义词词林扩展版》^②。

F8: 两个小句的句子相似度。F9: 两个小句的词汇最大相似度。

两个小句句子相似度越大,小句的关联程度越大。两个小句所含词的最大相似度越大,小句的关联程度越大。采用 Synonyms^③的 compare 方法得到句子相似度和词语相似度,比较两个小句中所有词语的相似度,找到最大相似度。

F10: 共现词(名动形副)出现个数。

共现词是指在同一篇章中以一定频率一起出现的词。当两个小句中的共现词越多,两个小句的关联程度就越大。从关联句对中提取名词、动词、形容词、副词,统计所有词之间分别在关联句对同时出现的次数 p,将 p 与这两个词在语料中出现次数积的比值作为一对词的分值 score,若 score 大于某个阈值,则视这对词为共现词。阈值是一个大于 0 小于 1 的数,调整时一步步缩小范围,计算并找到一个使得准确率

^① 本文均采用李艳翠等^[9]的小句切分标准。

^② http://www.ltp-cloud.com/download#down_cilin

^③ <https://github.com/huyingxi/Synonyms>

最高的阈值，最终确立为 0.001。

（三）句法

F11：句法结构上，两句相同词有共同祖先(不超过三层)的次数。**F12：**句法结构上，两句同义词有共同祖先(不超过三层)的次数。**F13：**句法结构上，两句同类词有共同祖先(不超过三层)的次数。

两个小句有相同词、同义词、同类词，且在句法结构上，其标记相同祖先节点(不超过三层)次数越多，关联程度越大。采用 `ctbparser`^① 对语料句法分析。如例 1 的小句对 3-6，出现了相同词“制定”，同时出现标记相同祖先节点(不超过三层)VP（分别为“制定法规条例”、“制定和推出法规性文件”的结构标记）。

（四）小句距离与句域

F14：两个小句的距离。**F15：**两个小句是否在同一个大句中；若在同一大句中其小句的数目。**F16：**特征 F14 与 F15 的组合。

两个小句距离越近，两个小句的关联的可能性越大。同属一个大句的小句，关联的可能性大于不同属一个大句的小句。同时，属于同一大句的小句对，该大句包含的小句数越少，大句中的小句对的关联程度越大；当一个大句有且仅有两个小句时，这两个小句必然关联。

2.2 实验数据

在汉语宾州树库^[8]的前 100 篇语料上，进行标注，得到共 2439 个小句，1956 个小句关联对，构成汉语篇章小句关联语料库，作为实验语料。实验采用 `sklearn`^② 对提取的特征进行分类，使用 10 次 10 折交叉的方法进行验证。一共提取了 6944 条小句对特征，其中有 4988 条负例，1956 条正例，正负比例 2:5。为了均衡正负例的比例，使用随机抽取正例的方法扩充数据集，使正负例比例为 1:1。

2.3 结果分析

2.3.1 分类器

使用决策树模型（`RandomForestClassifier`）、逻辑回归模型（`LogisticRegression`）、贝叶斯模型（`MultinomialNB`）三类分类器^③，采用以上全部特征。实验结果见表 1。

表 1 不同分类器实验结果

分类器	平均结果(%)			负例(%)			正例(%)		
	(准确率)	(召回率)	(F1)	(准确率)	(召回率)	(F1)	(准确率)	(召回率)	(F1)
<code>RandomForestClassifier</code>	92.70	92.62	92.61	94.61	90.38	92.45	90.80	94.85	92.78
<code>LogisticRegression</code>	87.36	87.34	87.34	88.25	86.16	87.19	86.48	88.53	87.49
<code>MultinomialNB</code>	84.66	84.65	84.65	85.30	83.73	84.51	84.03	85.58	84.79

由表 1 可知，分类模型效果：决策树模型>逻辑回归模型>贝叶斯模型。其中，决策树模型的准确率接近 93.00%。本文使用特征较多，且同义词、同类词等不同特征之间有一定关联。贝叶斯模型通常假设属性之间相互独立，在属性个数比较多，或者属性之间相关性较大时，贝叶斯的分类效果不好。逻辑回归模型擅长分析线性关系，特征空间很大时表现效果并不好，对极值比较敏感，容易受极端值的影响。而决策树对于较多特征且相关的情况处理则表现较好，且对大型数据源通常可以在相对较短时间得到良好结果。

2.3.2 去环与语料取样

小句关联结构从整体上为无环结构（见 1.3）。以上将小句关联看作独立个体，尚没有考虑小句关联结构的整体无环性。这里加入去环操作，即当整个段落中小句关联形成环，则从识别出的小句关联中，去掉关联可能性最低的，直至形成一个无环图。这里主要从语料独立/整体取样角度研究。

独立取样：使用 10 次 10 折交叉方法，对数据集随机抽取小句对构建训练集和测试集。整体取样：随机抽取单个段落进行划分，使用 10 次 10 折交叉方法，随机抽取正例对数据集进行扩充，使得数据集的所有小句对均能构成完整的段落。分别在独立取样语料和整体取样语料上进行去环实验，实验结果见表 2、3。

^① Paskin, Mark A. Cubic-time parsing and learning algorithms for grammatical bigram models. Computer Science Division, University of California, 2001.

^② <https://scikit-learn.org/stable/index.html>

^③ https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

表 2 独立取样与整体取样的去环结果

Table 2 De-ring results for independent and global sampling

分类器		独立取样									整体取样								
		平均结果(%)			负例(%)			正例(%)			平均结果(%)			负例(%)			正例(%)		
		(准确率	召回率	F1)	(准确率	召回率	F1)	(准确率	召回率	F1)	(准确率	召回率	F1)	(准确率	召回率	F1)	(准确率	召回率	F1)
RandomFore	去环前	92.70	92.62	92.61	94.61	90.38	92.45	90.80	94.85	92.78	80.48	78.64	78.31	72.99	90.93	80.98	87.98	66.35	75.65
stClassifier	去环后	92.27	92.21	92.21	93.82	90.38	92.07	90.72	94.04	92.35	85.51	81.14	80.55	73.06	98.68	83.96	97.97	63.61	77.14
LogisticRegr	去环前	87.36	87.34	87.34	88.25	86.16	87.19	86.48	88.53	87.49	82.63	82.45	82.42	80.16	86.23	83.09	85.11	78.66	81.76
ession	去环后	86.90	86.89	86.89	87.44	86.16	86.79	86.36	87.62	86.99	88.89	86.73	86.54	79.72	98.52	88.13	98.06	74.93	84.95
Multinomial	去环前	84.66	84.65	84.65	85.30	83.73	84.51	84.03	85.58	84.79	78.56	78.13	78.05	75.07	84.23	79.39	82.04	72.03	76.71
NB	去环后	84.28	84.28	84.28	84.65	83.73	84.19	83.91	84.82	84.36	86.20	82.95	82.56	75.35	97.94	85.17	97.05	67.96	79.94

表 3 独立取样与整体取样的去环结果的升降变化

Table 3: Ups and downs of de-ring results for independent and global sampling

分类器		平均结果变化(%)			负例变化(%)			正例变化(%)		
		(准确率 召回率 F1)	(准确率 召回率 F1)	(准确率 召回率 F1)	(准确率 召回率 F1)	(准确率 召回率 F1)	(准确率 召回率 F1)			
RandomForestClassifier	独立取样	-0.43	-0.41	-0.40	-0.79	0	-0.38	-0.08	-0.81	-0.43
	整体取样	5.03	2.50	2.24	0.07	7.75	2.98	9.99	-2.74	1.49
LogisticRegression	独立取样	-0.46	-0.45	-0.45	-0.81	0	-0.40	-0.12	-0.91	-0.50
	整体取样	6.26	4.28	4.12	-0.44	12.29	5.04	12.95	-3.73	3.19
MultinomialNB	独立取样	-0.38	-0.37	-0.37	-0.65	0	-0.32	-0.12	-0.76	-0.43
	整体取样	7.64	4.82	4.51	0.28	13.71	5.78	15.01	-4.07	3.23

根据表 2、3 可知，独立取样对去环影响不大，整体取样对去环影响较大。整体取样去环后，小句关联的正确率和召回率明显提高，尤其正例正确率和负例召回率提高较大。因此，小句关联识别中，需考虑整体取样。但需注意，整体取样去环后，最好识别效果与最佳分类器相比有变化，其中决策树性能降低，其他分类器性能变好，最好的为逻辑回归模型，平均准确率 88.89%。需考虑结合这些变化改进研究。

2.3.3 特征贡献

使用表 1 中最好的 RandomForestClassifier 进行实验，分别去除一类（组）特征后的实验结果见表 4。

表 4 特征贡献

Table 4 feature contribution

特征	平均结果（%）			负例（%）			正例（%）		
	(准确率	召回率	F1)	(准确率	召回率	F1)	(准确率	召回率	F1)
全部特征	92.70	92.62	92.61	94.61	90.38	92.45	90.80	94.85	92.78
-F1	92.60	92.49	92.49	94.71	90.02	92.30	90.49	94.97	92.68
-F2	92.67	92.59	92.59	94.46	90.49	92.43	90.87	94.70	92.74
-(F3+F4+F5+F6+F7)	92.70	92.60	92.60	94.72	90.24	92.42	90.68	94.97	92.77
-(F8+F9+F10)	87.83	87.74	87.73	89.64	85.35	87.44	86.02	90.13	88.03
-(F11+F12+F13)	92.88	92.79	92.79	94.78	90.58	92.63	90.98	95.01	92.95
-(F14+F15+F16)	88.00	87.80	87.78	90.81	84.11	87.33	85.20	91.49	88.23

根据表 4，对小句关联有较大影响的特征主要有：（1）两个小句所包含共现词的数量、两个小句的句子相似度及两个小句最大的词相似度。对于特征 F8、F9、F10，去除后整体准确率下降 4.87%，召回率下降 4.88%。（2）两个小句的距离，两个小句是否在同一个大句中及同一大句内小句的数目。对于特征 F14、F15、F16，去除后整体准确率下降 4.7%，召回率下降 4.82%。其他特征对小句关联识别的影响则不明显。

2.3.4 小句关联的距离

小句对距离对于判断两个小句之间是否有关联有重要影响。根据语料库统计（表 5），两个小句距离越近，相关的可能性越高；反之，距离越远，相关的可能性越低。其中相邻句对的相关比例近 70.00%。

表 5 小句对距离的相关/不相关分布

小句对 距离	相关			不相关			总计	
	数量	横比(%)	纵比(%)	数量	横比(%)	纵比(%)	数量	横比(%)
1	1352	69.44	69.12	595	30.56	11.93	1947	100.00
2	334	22.96	17.08	1121	77.04	22.47	1455	100.00
3	132	12.49	6.75	925	87.51	18.54	1057	100.00
4	50	6.61	2.56	706	93.39	14.15	756	100.00
5	32	5.99	16.36	502	94.01	10.06	534	100.00
5+	56	4.69	2.86	1139	93.31	22.83	1195	100.00
总计	1956		100.00	4988		100.00	6944	

在整体取样语料上，使用表 2 中表现最好的 LogisticRegression 分类器，区分相邻/不相邻小句对进行识别，识别结果见表 6。

表 6 相邻/不相邻小句对的识别

小句对 距离	平均结果(%)			不相关(%)			相关(%)		
	(准确率)	(召回率)	(F1)	(准确率)	(召回率)	(F1)	(准确率)	(召回率)	(F1)
相邻	93.24	90.80	91.48	62.75	91.88	74.57	98.48	90.61	94.38
不相邻	86.08	84.11	81.60	82.69	99.39	90.27	95.83	40.23	56.67

根据表 6：第一，不相邻小句对的识别难度大于相邻小句对，前者平均识别准确率较后者低 7.00%多。第二，无论小句是否相邻，难点均在不相关识别上，不相关识别比相关识别准确率平均低约 25.00%。值得注意，相邻小句对识别看似容易，但在不相关识别上，其识别准确率比相关识别低约 36.00%，甚至比不相邻小句对的不相关识别也低约 20.00%。小句对不相关识别，尤其是相邻小句对的不相关识别需要突破。

2.3.5 小句关联的句域

小句关联的句域，即小句是否在同一大句，对小句关联有影响。根据语料库统计（表 7），同一大句内的小句对相关的概率远大于不在同一大句内的小句内，前者是后者的 2 倍多。

表 7 在/不在同一大句的小句对的相关/不相关分布

小句对 句域	相关			不相关			总计	
	数量	横比 (%)	纵比 (%)	数量	横比 (%)	纵比 (%)	数量	横比 (%)
在同一大句	1350	46.26	69.02	1568	53.74	31.44	2918	100.00
不在同一大句	606	15.05	30.98	3420	84.95	68.56	4026	100.00
总计	1956		100.00	4988		100.00	6944	

在整体取样语料上，使用表 2 中表现最好的 LogisticRegression 分类器，对在/不在同一大句的小句对识别结果见表 8。

表 8 在/不在同一大句的小句对的识别

小句对 句域	平均结果(%)			不相关(%)			相关(%)		
	(准确率)	(召回率)	(F1)	(准确率)	(召回率)	(F1)	(准确率)	(召回率)	(F1)
在同一大句	90.71	88.41	88.74	74.12	96.73	83.93	98.27	84.61	90.93
不在同一大句	87.56	85.63	84.29	83.09	99.35	90.49	97.45	55.31	70.57

根据表 8: 第一, 不在同一大句的小句对识别难度大于在同一大句的小句对, 前者平均识别准确率低于后者 3.00% 多。第二, 无论小句是否在同一大句, 小句对不相关识别比相关识别准确率平均低近 20.00%。尤其是, 同一大句的小句对识别看似容易, 但在不相关识别上, 准确率比相关识别低约 24.00%, 甚至低于不在同一大句内的小句对不相关识别。小句对不相关, 尤其同一大句内的小句对不相关识别需研究突破。

3 结语

本文提出将篇章结构表示为小句关联结构, 相比篇章层次结构, 它可以有效刻画小句间的直接关联, 特别是可以刻画非连续和跨篇章层级的小句间的直接关联; 相比篇章依存结构, 它无需判断中心(但可据其推测中心), 单纯而更具基础性。提出小句关联的判断准则, 即直观上关联小句可以构成最小语篇, 理性上关联小句语义连贯、形式衔接。提出小句关联结构的形式限制, 即小句联通但无环, 及其语义解释。标注一致性实验表明, 所提的篇章小句关联结构具有较强可行性和科学性。

在自建的汉语篇章小句关联结构标注语料库上, 进行探测性小句关联识别研究。将小句关联识别作为判断任意小句有无关联的分类问题, 采用机器学习常见分类模型(决策树、逻辑回归、贝叶斯), 设计分类特征(连接词、相同词、同义词、同类词、共现词及有关句法结构相似度, 实词及句子相似度, 小句距离与句域等), 最佳识别准确率接近 93.00%。实验进一步发现: 语料整体取样较独立取样取得的效果好; 共现词数量、句子相似度、小句距离、小句句域等特征对小句关联识别贡献较大; 远距离和跨大句是小句关联分析的难点, 但相邻小句和同一大句内的小句对不相关识别则尤其难。

下一步我们将在本文小句关联表示基础上, 进一步扩大汉语篇章小句关联结构标注语料库规模, 并最终提供公开研究使用。同时, 我们也将改进小句关联的识别方案与计算模型等, 提高小句关联识别性能。并将研究基于小句关联的篇章中心识别等有关应用。

参考文献

- [1] Mann W, Thompson S. Rhetorical structure theory. Toward a functional theory of text organization. Text, 1988, 8(3):243-281
- [2] Carlson L, Marcu D, Okurowski M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory// Current and new directions in discourse and dialogue, Springer, 2003: 85-112
- [3] Prasad R, Dinesh N, Lee A, et al. The Penn Discourse Treebank 2.0//Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC). Marrakech, 2008:2961-2968
- [4] Li Yancui, Feng Wenhe, Kong Fang, et al. Building Chinese discourse corpus with connective-driven dependency tree structure//Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP). Doha, 2014:2105-2114
- [5] Li Sujian, Wang Liang, Cao Ziqiang, Li Wenjie.Text-level discourse dependency parsing//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore, 2014:25-35
- [6] Yang An, Li Sujian. SciDTB: Discourse Dependency TreeBank for Scientific Abstracts//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) . Melbourne, 2018: 444-449
- [7] Yoshida Y, Suzuki J, Hirao T, et al. Dependency-based discourse parser for single-document summarization//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, 2014:1834-1839
- [8] Hudson R A. Word grammar. Blackwell Oxford.1984
- [9] 李艳翠,冯文贺,周国栋,等. 基于逗号的汉语句子识别研究. 北京大学学报(自然科学版), 2013, 49(1):7-14
- [10] Xue N, Xia F, Chiou F D, et al. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 2005, 11(2):207-238