文章编号：1003-0077（2011）00-0000-00

基于汉英平行语料库的英文显式篇章关系识别 *

冯洪玉[1,2]，李艳翠[1,2]，冯文贺[3]，周国栋[2*]

（1. 河南科技学院信息工程学院，河南新乡 453003；2. 苏州大学计算机科学与技术学院，江苏苏州 215006；3. 河南科技学院文法学院，河南新乡 453003）

摘要：汉英篇章平行语料库有助于基于篇章的双语研究，该文构建了汉英平行语料库，对语料中的汉语及其英语对译中的连接词分别进行了标注和关系分类。其中英文连接词比单语语料上的英文连接词定义广泛，更为复杂。该文在此语料上，抽取词法、句法和位置信息等特征在英文文本上进行显式篇章关系识别，实验采用最大熵分类方法，获得连接词识别正确率 92.5%；抽取英文和对应中文连接词作为特征获得给定连接词关系分类正确率 85.6%。为今后的中英篇章关系对比识别提供参考。

关键词：显式篇章关系；连接词识别；分类；

中图分类号：TP391　　**文献标识码：**A

English Explicit Discourse Relation Recognition on Chinese-English Parallel Corpus

FENG Hongyu[1,2], LI Yancui[1,2], FENG Wenhe[3], ZHOU Guodong[2]

(1.School of Information Engineering, Henan Institute of Science and Technology, Xinxiang, Henan 453003, China; 2. School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China; 3. College of Humanities, Henan Institute of Science and Technology, Xinxiang 453003, China)

Abstract: Chinese-English discourse parallel corpus contributes to bilingual discourse research. This paper constructed the Chinese-English discourse Parallel Corpus, which annotates conjunctions and relation classification in Chinese corpus and English corpus. In this corpus English conjunction definition is wider than traditional conjunction's, and is more complicate. On this corpus, the paper extracts lexical, syntactic features and location information to identify and classify the explicit discourse relation in the English text. Experiment adopts with maximum entropy classification method to obtain conjunction recognition accuracy of 92.5%; and extracts English and Chinese conjunction as features to obtain given conjunction classification accuracy of 85.6%. The results provide a reference for contrast recognition of Chinese-English discourse relation for future.

Key words: explicit discourse relation；conjunction recognition；classification

1 引言

篇章关系是指同一篇章内部，句子与句子或子句与子句之间的语义连接关系[1]，根据是否有连接词，篇章关系分为显式关系和隐式关系。近年来, 由于大规模人工标注的英文篇章语料库的出现，英语篇章关系识别的研究颇多，使用较多的语料库如修辞结构理论篇章树库(Rhetorical Structure Theory-Discourse Treebank, RST-DT)[2]和宾州篇章树库(Penn Discourse TreeBank, PDTB)[3]等。汉语篇章语料目前主要有汉语复句语料库、清华汉语树库、哈工大中文篇章关系语料库，文献[4]标注了基于连接依存树的汉语篇章结构语料库。

虽然单语篇章语料库都有了一定的发展，但对汉英篇章结构平行语料库的研究还较少。平行语料库为基于篇章的互译技术的发展提供支持。它的标注规范和单语语料上的标注会有一定的区别，包括连接词的定义和关系分类等。

收稿日期：2015-6-15 **定稿日期：**2015-8-10

基金项目：国家自然科学基金（61273320）；教育部人文社科项目(13YJC740022)；河南省教育厅科学技术研究重点项目(14A520080)

本文标注了汉英篇章平行语料，语料选取 OntoNotes 中的新华日报及其对译文本，目前标注了 100 篇中英对译文本（chtb0001-0175 中的 100 篇）。该汉英平行语料以汉语为主，英语为指导的原则进行标注。由于该语料是汉英方向的平行语料，连接词定义和关系分类按汉语分类方法制定，英文语料中标注的英文连接词有别于传统的英文连接词，定义更为广泛。除了常规的英文连接词以外，有些非连接词如“of which”，“among which”等都作为了连接词，它们是由中文连接词“其中”翻译过来的，这些连接词一般在英语单语语料中（如 PDTB）作为普通词汇。所以本文识别的连接词相对复杂。基于汉英平行语料库的篇章关系识别对于基于篇章的互译技术是很重要的，目前相关文献中还未见到有对汉英平行语料中的英文连接词进行识别的工作。本文抽取词法、句法和位置信息等特征在英文文本上进行显式篇章关系的识别。并且在分类识别时加入中文连接词作为特征，提高了分类准确率。最后本文还将识别结果和文献[5]在对应的中文语料上的识别结果进行对比，为今后的中英篇章关系对比识别提供参考。

2 相关工作

现有英文语料资源主要包括修辞结构篇章树库（RST-DT）和宾州篇章树库（PDTB）等。RST-DT 由 LDC（Linguistic Data Consortium）于 2002 年发布。RST-DT 摘取宾州树库中的 385 篇华尔街日报进行标注，篇幅长度不等，平均每篇文章 458 个词。文章的内容涉及到各种话题，其篇章结构分析的首要任务是确定基本篇章单位（Elementary Discourse Unit，EDU）。EDU 确定下来后，再根据 RST 确定基本篇章单位之间的关系，进而生成有层次的篇章结构树。

PDTB 专注于描述篇章关系，是目前规模最大的英文篇章级别的语料库。由美国宾西法尼亚大学、意大利托里诺大学和英国爱丁堡大学联合标注，于 2008 年由 LDC 发布。主要标注与篇章连接词相关的连贯性关系。标注信息主要包括篇章连接词的论元结构、语义区分信息，以及连接词和论元的属性相关特征等。主要标注关系有：显式关系（Explicit）、隐式关系(Implicit)、替代关系（AltLex）、实体关系（EntRel）、无关系（NoRel）。并为每种论元结构关系标注了语义区分信息和属性信息。其中显式关系、隐式关系、替代关系都标有语义区分信息和属性信息，EntRel 和 NoRel 没有标注。语义结构分为三级层次：种类—类型—子类型。第一层包括 Temporal、Contingency、Comparison 和 Expansion 在内的 4 类语义，第二层包括 16 类语义，第三层包括 23 类语义。

在篇章关系识别方面，文献[6]在 PDTB 上进行显示篇章关系识别，文中提到采用连接词作为特征取得了 93%的连接词识别准确率。进而采用连接词本身、词性以及句法作为特征取得了 96.26%的连接词识别准确率和 94.15%的关系类别识别准确率。

中文篇章分析的不足主要在于一是中文篇章理论的不成熟，二是大规模中文篇章语料库的缺乏。中文篇章语料库方面的工作主要集中在以下两个方面：以句群和复句理论为代表的中文篇章语料库；借鉴西方 RST 和 PDTB 体系的中文篇章语料库。目前包含连接词标记的汉语语料库主要有汉语复句语料库、清华汉语树库、哈工大中文篇章关系语料。文献[7]利用词性标记和关系词搭配理论进行关系词提取，提取的正确率达到 89.8%，对连用关系标记标识准确率达 72.9％。

清华汉语树库（Tsinghua Chinese Treebank，TCT）标出了复句内各分句之间的关系信息，但没有标注特定复句关系所对应的复句关系词。文献[8]利用规则从 TCT 中提取复句关系词并标注其类别，抽取的句法、词法、位置特征进行篇章关系词的识别和分类，实验结果表明关系词识别准确率达 95.7%，篇章关系识别的 F1 值为 77.2%。

哈工大中文篇章关系语料标注采用宾州篇章树库的模式。标注的关联词分为显式关联词和隐式关联词两种。篇章关系共分为 6 个大类：时序关系、因果关系、条件关系、比较关系、扩展关系和并列关系。

综上可知，目前单语语料上的标注工作相对较多，中文语料的连接词标注和分类也多参考英语的分类方法，汉英平行语料方面的工作还很鲜见，目前篇章关系识别也仅在单语语料上进行。本文以中文篇章关系分类规则为主，标注了汉英篇章结构平行语料库，该语料标注的英语连接词比传统的单语语料连接词复杂。基于该汉英对齐语料本文在英文语料上进行显式篇章关系识别。

3 汉英篇章结构平行语料库与连接词

3.1 汉英篇章结构平行语料库

汉英篇章结构平行语料库（Chinese–English Discourse Treebank,简写为 CEDT）是为汉英双语翻译文本标注了对齐篇章结构信息的语料库 [9]。如上所述，目前基于单语的篇章结构语料库所做工作较多，双语对齐篇章结构知识资源还相当匮乏，这也直接制约了基于其上的篇章结构机器翻译等研究的发展 [9]。本文实验采用的语料是 OntoNotes 中的新华日报及其对译文本，目前标注了其中 100 篇中英对译文本（chtb0001-0175 中的 100 篇）。本文对语料中的汉语及其英语对译中的连接词分别进行了标注和关系分类。该汉英平行语料以汉语为主，英语为指导的原则进行标注。英文连接词定义和关系分类按汉语分类方法制定，因此出现的英文连接词有别于常规的英文连接词，定义更为广泛。本文下面将其中标注的中文语料简称为 CDTB(Chinese Discourse Treebank)，其标注规范见文献[4]，对应的英文语料简称为 EDTB(English Discourse Treebank)。下面只介绍和本文有关的篇章关系标注情况。例 1a 和例 1b 是本文篇章结构平行语料库中的一个标注实例，基于结构对齐。带有下划线的为连接词，“|”表示第一层篇章关系划分。“||”表示第二层，以此类推。表 1 是例 1a 和例 1b 对应的部分具体标注结果。

例 1a：**为**实现上述规划，|沙头角保税区除继续完备相应的物质条件和政策法规、管理体制、运作机制、人才等条件外，||**并**推出下列措施予以保障。

例 1b:**To** accomplish the above program ,| the Shatoujiao Bonded Area not only continues to perfect its relevant material conditions and conditions such as policy regulation , management system , operating mechanism , qualified personnel , etc. ,|| **but also** put forward the following measures as a guarantee .

表 1 汉英篇章结构平行语料库标注实例

<ID="2">
RID="1" StructureType="逐层切分" ConnectiveType="显式关系" Layer="1" RelationNumber="单个关系" Connective="为" RelationType="目的关系" ConnectivePosition="1 … 1" ConnectiveAttribute="不可删除" RoleLocation="normal" LanguageSense="true" Sentence="为实现上述规划，|沙头角保税区除继续完备相应的物质条件和政策法规、管理体制、运作机制、人才等条件外，并推出下列措施予以保障。" SentencePosition="1…8|9…61" Center="2" ChildList="2" ParentId="-1" UseTime="21" />

<ID="2">
<RID="1" StructureType="逐层切分" ConnectiveType="显式关系" Layer="1" RelationNumber="单个关系" Connective="to" RelationType="目的关系" ConnectivePosition="" ConnectiveAttribute="不可添加" RoleLocation="normal" LanguageSense="true" Sentence="To accomplish the above program ,| the Shatoujiao Bonded Area not only continues to perfect its relevant material conditions and conditions such as policy regulation , management system , operating mechanism , qualified personnel , etc. , but also put forward the following measures as a guarantee ." SentencePosition="1 … 33|35 … 298" Center="2" ChildList="2" ParentId="-1" UseTime="7" />

CEDT 的核心思想是结构对齐，标注基本原则是“结构对齐，关系对齐”[9]。该语料的中英文标注体系是一致的，汉英文本的内部层次结构和关系一一对应。如例 1a 划分了两个层次，例 1b 也划分了两个层次，且与例 1a 结构对应一致。例 1a 中连接词“为”和“并”对应例 1b 中的“to”和“not only…but also”。但有时汉英连接词不是同时出现的，比较多的情况是，汉语中没有显式连接词，而英语中需要翻译出连接词。

3.2 连接词

由于英语语法特点，英文连接词一般比较集中和明显，如表示时间的“when”，“befor”，“since”，“at last”等，表示因果的“because”，“therefore”，“as a result”，“thus”等等。根据汉语语法特点，篇章连接词构成比较复杂。汉语中连接词主要指连接子句与子句，表示并列、选择、递进、转折、条件、因果等语法关系，表示连接作用的连词、关联词以及其他与之有同等关系作用的语言单位。由于本文处理的连接词是汉英对译下英语连接词，必然存在其特殊性，定义也同汉语一样较为广泛，只要对句子、子句或语段起连接作用，能正确表示语言单位之间关系的词语均可称为连接词。比如“among which”，“of which”，“among these”，“in this”都是连接词，如例 2。这与传统意义的英文连接词不一样。而且有些连接词并不是连接句子，可能连接的是一个短语，如例 1b 中的“to”，连接的是一个短语结构。这就造成本文处理的英语连接词可能会有数据稀疏问题。因此本文在分类识别时加入中文连接词作为特征，以达到提高识别率的目的。而且汉语中没有连接词，但翻译时出现英语连接词的情况也很常见。本文抽取了连接词 64 个，其中将联合连接词划分为 2 个连接词处理，如“not only…but also”和“both…and”之类的连接词。

例 2：So far, there are already 410 enterprises in the whole zone, **among which** 223 have been identified as new, high level technology enterprises.

EDTB 中共标注显式关系 462 次。出现次数最多的 10 个连接词及次数如表 2。出现次数较多的前 5 个连接词占 74.9%，只出现 1 次的连接词有 43 个，占 9.3%。

表 2 连接词出现次数排序表（前 10）

连接词	次数	连接词	次数
and	264	in order to	11
to	35	of which	10
at the same time	17	moreover	6
but	16	among them	5
among which	14	so as to	4

连接词在连接篇章单元的同时，也表示它们之间的语义关系。CEDT 参考文献[10]的分类方法，将篇章关系分成三个意义层次：第一层 4 大类：因果类、并列类、转折类、解说类；四大类下面细分为第二层，共 17 个小类，第三层为连接词，具体关系分类层次见表 3。

表 3 篇章关系分类层次表

第一层	第二层	第三层
因果类	因果关系、推断关系、假设关系、目的关系、条件关系、背景关系	thus、to、as 等
并列类	并列关系、顺承关系、递进关系、对比关系、选择关系	and、both…and、while 等
转折类	转折关系、让步关系	however、but、instead 等
解说类	解说关系、总分关系、例证关系、评价关系	all this shows 等

在篇章结构关系中，连接词存在两种类型的歧义问题需要处理。一种是对某词是连接词

还是非连接词的判断。最明显的如“and”，“and”可以作为并列类连接词，也可以作为普通词汇，例 3 中“and”代表具体含义“和”，并不是连接词。再一种歧义是连接词和关系类型并不是一一对应的关系。一个关系类别有很多连接词，如：因果类别中可以包含“because”，“so”，“as a result”，“and therefore”等连接词；一个连接词也可以表示多个关系类别，如在已标注的文档中，以出现次数最多的连接词“and”为例，连接词“and”可以表示 5 种关系，分别是“并列关系”出现 256 次，顺承关系出现 4 次，递进关系出现 1 次，因果关系出现 2 次，总分关系出现 1 次，具体参见例 3 至例 8。连接词识别就是确定该连接词是否具有篇章连接作用，篇章关系分类主要任务是根据连接词及其上下文特征判别连接词的语义类别。本文只进行第一层四大类关系的识别。

例 3：Because of this, new situations **and** new questions that have not been encountered before are emerging in great numbers.（不是连接词，和）

例 4：An office of Shanghai Customs posted at Chongming, that was approved by the China Customs Head Office to be set up, was established a few days ago, **and** has already officially conducted business.（并列类，并列关系）

例 5：The technological content is relatively high, **and** 60% of the enterprises possess the resources for high - tech products.（解说类，总分关系）

例 6：Concerned departments of the Henan provincial government promulgated this province's foreign economy and technological cooperation projects at the conference, **and** conference representatives negotiated on the cooperative intention of related projects.（并列类，顺承关系）

例 7：They make diligently study advanced technologies and modern management experience in foreign invested enterprises, **and** from one aspect, brings along the increase in citizen's quality.（因果类，因果关系）

例 8：The economy and trade exchange between Tianjin, an important economic city of north China , and the Russian Federation is currently steadily developing, **and** has shown new features.（并列类，递进关系）

4 实验与分析

本节对平行语料库中的英文语料显式篇章关系进行识别。实验内容包括连接词识别和给定连接词的关系分类识别。参与训练的文件 100 篇，段落数 413 段，篇章关系共 1492 条，其中显式关系 462 条，占 31%。为了充分利用语料资源，实验采用 10 倍交叉验证的方式，抽取词法、句法和位置特征。句法树利用斯坦福句法分析工具生成自动句法树。连接词识别是二元分类问题，篇章关系识别是多元分类问题，分类方法使用 Mallet[11]工具包中的最大熵模型。

4.1 篇章连接词识别

4.1.1 实验

例 3 句子的句法树如图 1 所示,例 4 句子的句法树如图 2 所示。句法树中与本文实验无关的句法部分用“…”代替。

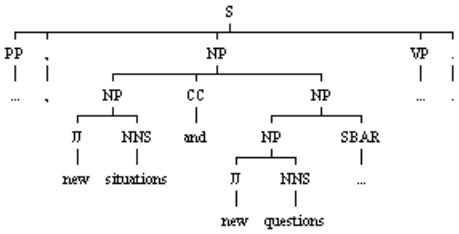

图 1 例 3 的句法树

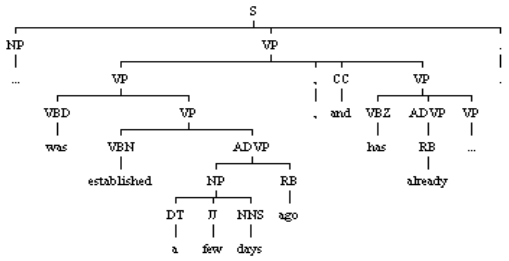

图 2 例 4 的句法树

实验选取特征主要参考文献[8]，但是文献[8]对应的是汉语语料的篇章连接词识别，和英文连接词会有一定的差别，根据英语语言现象，对某些特征进行调整，如英语连接词中经常出现多词连接词，如“so as to”，“at the same time”等，而汉语不存在这种现象。所以本文对多词连接词的词性特征选取为父节点信息。并且关系词左右兄弟信息可以反映连接词在句中的位置，所以连接词在句子中的位置特征不再提取。具体特征说明如下。

词法和语法特征：词法特征包括关系连接词本身和词性。如例 4 中为“and”和“cc”；对于多词连接词，词性取关系词的父节点信息，如“so as to”，(ADVP (RB so) (RB as) (S (VP (TO to)，则词性特征为“ADVP”。

句法特征：关系词节点的父节点信息和左右兄弟节点信息。如例 3“and”不是连接词时，父节点信息为“NP”，左右兄弟也为“NP”。例 4 中“and”是连接词，父节点为“VP”，左兄弟节点为“，”，右兄弟节点为“VP”。当连接词位于句首时，左兄弟记为“NONE”，同样，位于句尾时，右兄弟记为“NONE”。

位置特征：关系词是否位于句首和关系词前是否有标点。连接词常常会出现在子句的句首位置，如例 4 中的“and”。

表 4 列举了例 3 和例 4 中对“and”提取的所有特征，表 5 给出了连接词识别的实验结果。

表 4 特征表

	特征	例 3 特征值	例 4 特征值
词法	词本身和词性	and, cc	and, cc
	关系词前后 2 个词及词性	(new JJ) (situations NNS) (new JJ) (questions NNS)	（ago RB) (, ,) （has VBZ) (already RB)
句法	父节点	NP	VP
	左兄弟	NP	,
	右兄弟	NP	VP
位置	关系词是否位于句首	否	否
	关系词前是否有标点	否	是

表 5 自动句法树下关系词识别结果

语料	特征	正确率

EDTB 语料	词法	92.1
	词法+句法	92.4
	词法+句法+位置	92.5
CDTB 语料	词法+句法+位置	87.2[5]

4.1.2 分析

从表 5 可以看出，仅使用词法特征可以取得 92.1%的正确率，使用词法、句法和位置特征取得了 92.5%的正确率。中英语料使用相同的特征，本文训练了 100 篇英文语料，文献[5]中使用了 500 篇中文语料，连接词识别正确率 87.2%，可以看到英文连接词识别准确率明显高于中文。通过语料比对，100 篇英文语料中，英文显式关系 462 条，对应到 100 篇中文语料中只有 242 条。说明汉语中没有显式连接词，而英文翻译出了连接词的情况是很常见的。但是本文比文献[6]中 P&N 测试的 PDTB 语料库连接词识别准确率低，原因是 PDTB 语料中的连接词都是常规的英语连接词，在文中出现频率较高，比较集中，识别效果相对较好。再者本文使用的语料库规模较小，P&N 处理了 18459 条实例，包含 100 个连接词 [6]，本文处理实例个数为 5057 条，包含 64 个连接词。而 64 个连接词中只出现 1 次的连接词有 43 个，占了 67.2%。

4.2 连接词分类识别

4.2.1 实验

经统计，语料中共标注显式关系 462 条，其中 4 大类关系的分布如表 6 所示，可以看到，仅并列关系有 332 条，占一半以上，转折关系所占比例最少。鉴于语料规模还较小，连接词存在数据稀疏问题，本次实验仅对 462 条给定的关系进行分类，由于本文所用语料是汉英平行语料，特征提取了英文连接词本身和对应的中文连接词两项特征，关系类别的识别对连接词本身的依赖性很强，使用英汉连接词本身可以取得较好的分类效果，表 7 给出了连接词作为特征的分类性能。

表 6 关系类别的分布

关系类别	出现的次数	所占比例%
并列类	332	71.86
因果类	64	13.85
解说类	43	9.31
转折类	23	4.98

表 7 给定连接词的识别结果

关系类别	正确率	召回率	F1 值
并列类	95.5	90.0	92.6
因果类	94.4	77.4	83.5
转折类	30.0	88.1	44.3
解说类	95.8	67.0	72.8
平均	85.6		

4.2.2 分析

从表 7 可以看到，并列类，因果类和解说类分类性能较好，转折类识别效果较差。因为并列类训练实例较多，连接词比较集中，所以识别率比较高。但是由于存在一个连接词对应多种关系的情况，如“and”在并列类中出现 261 次，解说类中出现 1 次，因果类中出现 2 次，所以导致一些识别错误。解说类关系虽然所占比例较低，在语料中共出现 43 次，但是

解说类连接词比较明显，如出现时多集中为“among”，“among which”，“of which”等词语，所以解说类识别准确率较高。

转折类准确率低，一是因为语料中转折类关系数量少，总共只出现 23 次，连接词共有 7 个，分布情况是：“but”出现 15 次，“however”出现 3 次，“nevertheless”，“yet”，“but at the time”，“but now”，“as long as”均只出现一次。数据稀疏，造成转折关系识别不准确。二是因为有些转折类连接词还对应了其它的关系类别，如“but”在转折类中出现 15 次，在并列类中出现 1 次，而分类时一般会将之归为标注次数较多的一类，本次实验将例 9 中的 but 错误地归为了转折类。

例 9：According to investigations , the annual global sale of herbal medicine is about 15 billion US dollars ,**but** China 's export of Chinese medicine is only 600 million US dollars , and 70 % of the medicinal materials are have no added value .（英 chtb_0057, 并列类、对比关系）

经实验分析得知，同一连接词对应的关系类别越少，该词的歧义性越小，每个类别的连接词越集中，出现频率越高，连接词类别识别率越好。以后通过增加语料规模，训练结果会达到实用效果。

5 结论

本文选取 OntoNotes 中的新华日报及其对译文本进行篇章级别对齐标注，在标注的 100 篇英文语料上实现显式篇章关系的识别。该语料是汉英方向的平行语料，连接词定义和关系分类按汉语分类方法制定，因此英文语料中出现的连接词相对复杂，有别于传统的英文连接词。该文根据语料特点，给出了平行语料下英文连接词识别结果和实验分析，从统计和实验结果看，英文语料比对应的中文语料连接词识别率要高，下一步可以扩大平行语料库规模，进一步挖掘汉英对译关系的特点，为中英连接词对比识别打下基础性工作，进而为汉英机器翻译提供支持。

参考文献

[1] 周小佩，洪宇，车婷婷，等. 一种无指导的隐式篇章关系推理方法研究[J]. 中文信息学报，2013, 27(2):17-25.

[2] Carlson L, Marcu D, Okurowski M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory[C]. Proceedings of the SIGDIAL. Stroudsburg: Association for Computational Linguiscs, 2001: 110

[3] Prasad R, Dinesh N, Lee A, et al. The Penn Discourse Treebank 2.0[C]. Proceedings of the 6th International Conference on Language Resources and Evaluation.2008：2961-2968

[4] Li YC, FengWH, Sung J et al. Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure[C]// Proceedings of EMNLP2014. 2014:2105-2114.

[5] 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). 2015,2: 307-314

[6] Pitler E, Nenkova A. Using syntax to disambiguate explicit discourse connectives in text[C]//Proceedings of the ACL-IJCNLP 2009. Stroudsburg: Association for Computational Linguistics, 2009: 13-16

[7] 胡金柱，舒江波，姚双云，等. 面向中文信息处理的复句关系词提取算法研究[J]. 计算机工程与科学, 2009, 31(10):90-93

[8] 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.

[9] 冯文贺.汉英篇章结构平行语料库的对齐标注研究[J].中文信息学报,2013(6):158-165.

[10] 黄伯荣，彦序东. 现代汉语（下册）[M]. 北京：高等教育出版社，2002.

[11] McCallum A K. Mallet: a machine learning for language toolkit [CP/OL]. (2002)[2012.2.28].

http://mallet.cs.umass.edu

作者简介：

	冯洪玉(1977－)，女，讲师，硕士，主要研究领域为自然语言处理。 Email：feng_hongyu@126.com		李艳翠(1982－)，女，讲师，博士研究生，主要研究领域为自然语言处理。 Email: yancuili@gmail.com
	冯文贺(1977－)，男，讲师，博士，主要研究领域为计算语言学。 Email：Wenhefeng@gmail.com		周国栋(1967－)，男，教授，博士，博士生导师，主要研究领域为自然语言处理、多语言跨文本信息抽取。通讯作者,Email: gdzhou@suda.edu.cn