汉英篇章衔接对齐语料库构建研究

李艳翠*　冯继克　来纯晓　冯洪玉
河南科技学院信息工程学院，新乡 453003
liyancui@hist.edu.cn

摘要

篇章衔接性分析是理解篇章的基础，汉语和英语在指代、连接和省略等主要衔接方式上存在差异。本文旨在创建汉英篇章衔接对齐语料库，给出包括子句、连接词、指代和省略的汉英篇章衔接对齐标注策略，建立包含相应信息的对齐信息的语料库资源，最后对标注语料进行评估并讨论了标注中的难点问题及解决方法。对语料库标注质量评估及简单实验结果表明，本文研究语料标注策略方法切实可行，所标注的资源一致性满足实际需要。

关键词：篇章衔接；对齐语料标注；指代；省略；连接

Research on the Construction of Chinese-English Discourse Cohesion Alignment Corpus

Yancui Li　Jike Feng　Chunxiao Lai　Hongyu Feng
College of Information Engineering, Henan Institute of Science and Technology, Xinxiang 453003
liyancui@hist.edu.cn

Abstract

Discourse cohesion analysis plays a critical role in discourse understanding, in which there exist differences in cohesion between English and Chinese, including anaphor, ellipsis and connective. This paper aims to create a corpus containing corresponding cohesion alignment information. First, we explore proper strategies in annotating discourse cohesion, including clause, conjunction, reference and ellipsis. Then, we create resources which contains the information of alignment. Finally, this paper evaluates the corpus, discusses the problems and solutions in the annotation. The evaluation of corpus labeling and simple experimental results show that the method of corpus labeling strategy in this paper is feasible and the consistency of labeled resources meets the actual needs.

Keywords: Discourse Cohesion; Alignment Corpus Annotation; Anaphor; Ellipsis; Conjunction

1 引言

自然语言的单位可以从小到大分为词、短语和句子，最后形成一个篇章。在实际应用中，自然语言处理大都要在篇章上进行，不可断章取义，要正确理解篇章，就需要了解篇章中的衔接。衔接是一个语义概念，当篇章中的某个成分的含义需要依赖于另一个成分解释时，就会出

现衔接，汉语和英语中都有多种衔接手段。衔接主要有指代、省略和连接：指代是指用代词、冠词等表示特定的事物或已被提及过的事件；省略是指在事理逻辑上应有在字面上却没有的成份；连接主要指连接不同篇章并表达语义关系（例如因果、并列、转折等）的词语。汉英篇章衔接手段有差异，如例 1 和例 2。

例 1a: (他)^{r1} 脱下衣服的时候 c1，他 ^{a1} 听得外面很热闹，阿 Q^{a2} 生平本来最爱看热闹，(他)^{r2} 便 ^{c2} 即寻声走出去了。(他)^{r3} 寻声渐渐的寻到赵太爷的内院里，虽然 ^{c3} 在昏黄中，(他)^{r4} 却 ^{c4} 辨得出许多人，赵府一家 ^{a3} 连两日不吃饭的太太也在内，还有 ^{c5}(他们)^{r5} 间壁的邹七嫂，(也有)^{c6}(他们)^{r6} 真正本家的赵白眼，赵司晨。 (鲁迅：阿 Q 正传)

例 1b: **While**$^{c1'}$ **he**$^{r1'}$ was taking off his shirt **he**$^{a1'}$ heard uproar outside, **and since**$^{c2'}$ **Ah Q**$^{a2'}$ always liked to join in any excitement that was going, **he**$^{r3'}$ went out in search of the sound, **he**$^{r4'}$ traced it gradually right into Mr. Chao's inner courtyard. **Although** $^{c3'}$ it was dusk **he**$^{r4'}$ could see many people there: **all the Chao family** $^{a3'}$ including the mistress who had not eaten for two days. **In addition** $^{c5'}$, **their**$^{r5'}$ neighbor Mrs. Tsou was there, **as well as** $^{c6'}$ **their** $^{r6'}$ relatives Chao Pai-yen and Chao Szu-chen. (杨宪益、戴乃迭译：The True Story of Ah Q)

例 2a：尽管 ^{c1} 减轻污染 ^{a1} 的呼声不断，(并且)^{c2} 公众日渐愤怒，污染 ^{a2} 还是变得更糟糕了，(这)^{r1} 越发显出环保的紧迫性。

例 2b:**Despite**$^{c1'}$frequent calls for cutting **pollution**$^{a1'}$, **and**$^{c2'}$ growing public anger, **the problem**$^{a2'}$ has only got worse, **which**$^{r1'}$ increasingly shows the urgency of environmental protection.

例 1 中的篇章衔接方式主要有指代、省略和连接。例 1a 省略了四个主语"他"(r1-r4)，由于省略的主语在上下文中是暗含的，因此并未给读者在阅读上造成困难，省略的"他"和阿 Q 形成省略衔接；但在同样的情况下，如例 1a 的对照翻译例 1b，在英语中，主语是不能省略的，否则句子的结构上将不完整，翻译时被省略的主语 he(r1'-r4')都补充上。例 1a 中的他和阿 Q 和例 1b 中的 He 和 Ah Q 形成指代衔接。例 1a 中的连接成分"虽然"(c3)、"还有"(c5)、"也有"(c6)分别和例 1b 中的"Although"(c3')、"In addition"(c5')、"as well as"(c6')相对应，它们的功能相同，其中，连接词"也有"(c6)在汉语中是省略的，而相应的翻译中却根据意义补充了"as well as"(c6')。例 1 给出的例子反映了汉英衔接的实际情况，例 2 是(Tu Mei Tu Mei et al. ,2014)文中的实例，在翻译时，连接词"尽管"(c1)和"Despite"(c1)相应，"污染"(a2)在翻译时变成了"the problem"。综合分析例 1 和例 2 可知，汉英篇章中都存在各种衔接，衔接手段略有差异。

本文开展的汉英篇章衔接研究具有非常重要的理论意义和应用价值，形成的汉英篇章衔接对齐标注策略可用于构建语料库，所构建的语料库既可用于汉英篇章衔接的对比、翻译、教学等研究，又有助于推动汉英篇章衔接对齐分析研究及平台建设。

2 相关工作

2.1 汉英篇章衔接理论研究

Halliday and Hasan(1976)、Werth(1984)和 Cook(1989)分别将衔接进行了分类，文章中均指出主要的衔接手段包括连接、省略和指代。胡壮麟(1994)在《语篇的衔接与连贯》第一次系统地介绍了汉语篇章衔接与连贯，这本书是胡壮麟先生对 Halliday and Hasan(1976)衔接理论的继承和发展，除了保留 Halliday and Hasan(1976) 以语法和词汇为重点的衔接模式外，该书还包含了英语和汉语实例，这对汉英篇章衔接的研究具有很大的启示。周利芳(2018）和曹继阳(2019）分别对汉语篇章衔接的成分和手段进行了研究和分析。理论研究方面，汉英语篇的衔接基本都包括指代、省略、连接等，汉英语篇的衔接对比也多从这几个方面展开。奚雪峰等(2019）从篇章意图性角度探讨了篇章话题结构，并在此基础上分析了篇章的连贯性和衔接性。朱永生等(2001)的《英汉语篇衔接手段对比研究》将衔接理论用于汉英篇章对比，该书基于 Halliday and Hasan(1976)的衔接理论，运用大量的语料分析了英汉衔接手段的异同。由于汉语是一种意合型语言，人们在选择词语和句子方面通常能省则省，英语中大多数的省略都带有形式上的标记，

而汉语的省略是在不用考虑语法，甚至不用考虑逻辑的情况下表达其含义。此后许多研究者将衔接理论用于汉英语篇对比研究（常阳,2015; 钟书能,2016; 张献丽,2017;王菲,2018; 张易男和李燕鸿,2019），这些论文大多数采用 Halliday and Hasan(1976)对衔接手段的分类结合汉英语料分析汉英篇章衔接方式的异同。以上汉英对比研究取得了一定的效果，但选择的样本均较少，往往难以排除随机性对结果的影响。英汉对比研究应着眼于两种语言的特色，各自不同的趋势，需要选择有代表性且较多的样本。

2.2 汉英衔接语料库

语料库在自然语言处理技术的发展过程中起到了非常重要的作用。下面介绍包含指代、连接信息标注的语料库，以及汉英平行语料库。

包含指代信息的语料库。目前较知名的标注了指代信息的语料库主要有 MUC(Message Understanding Conference) 、ACE(Automatic Content Extraction) 、OntoNotes 语料库。MUC 语料通过指向形成指代链。ACE 中具有相同指代关系的实体位于同一指代链，且该指代链拥有唯一的编号。但 MUC 和 ACE 只标注了实体指代，并且没有考虑省略的指代标注。OntoNotes 包括词汇层面，句子层面和篇章层面多层次的标注，在篇章层面主要包含空语类信息、实体间以及事件的共指关系(Pradhan et al.，2007)。OntoNotes 中包含汉语和英语，汉语部分还标注了部分零指代信息，但零指代仅标注了主语位置，而汉语的零指代种类很多，且每一类别都有其自身的特点，这就制约了汉语零指代消解的研究。Kong and Zhou(2010)在 CTB6.0 语料标注的空语类(Empty Category)基础上进行了汉语零指代信息的标注，该语料有 150 篇文本。

包含连接信息的语料库。包含连接信息的语料库主要有宾州篇章树库(Penn Discourse Tree Bank)、汉语复句语料库、清华汉语树库、哈工大中文篇章关系语料以及苏州大学和河南科技学院合作完成的汉语篇章结构语料库（CDTB）。以上对于篇章的标注多采用英语篇章体系，李艳翠等(2015)提出一种基于连接依存树的汉语篇章结构表示方法，连接依存树的主要特征是叶子节点为子句，内部节点为连接词，连接词通过其层级地位(管辖范围)表示篇章结构的层次，通过其语义(具体与抽象)表示篇章关系。在此基础上，作者标注了 500 个文档的汉语篇章语料，其中有 24.8%的篇章关系有显式连接词。以上语料中虽然都涉及了连接词的相关标注，但均针对单语，篇章关系中汉语仅 25%左右有连接词，英语有 45.5%，可见英语连接词使用频率大于汉语。

平行语料库主要是指语料中的两种语言文本构成互译关系，目前的汉英衔接语料库主要针对单语，现有的汉英平行语料库除了做了一般性段落、句子等对齐工作外，很少进行语义等深度加工，特别是篇章层面的标注加工。因此，很难利用现有平行语料库进行基于篇章衔接的自动分析和应用研究。

综上，由于汉英衔接理论不同，衔接方式也有差别，汉英衔接对比多从指代、省略和连接方面进行，但目前的对比选择的样本均较少，不具有统计学意义。目前的汉英衔接语料库主要针对单语，现有的平行语料库只做了段落、句子等对齐工作，很少进行篇章衔接等深度加工，特别是衔接信息的对齐。这严重制约了基于篇章衔接对齐语料的语言对比及自动对齐分析工作。

3 汉英篇章衔接对齐标注策略

在充分分析现有汉英衔接理论、衔接对比分析理论和汉英衔接自动分析研究内容的基础上，本文制定了标注策略。词汇衔接由于有明显的词语指示，不是汉英衔接研究的难点，所以本文重点标注语法衔接，包括指代（本文将衔接理论中的指称和替代合并为指代）、连接和省略信息。杨传鸣（2008）对红楼梦及其英译本的衔接进行定量统计，发现在所有衔接手段中(包括词汇衔接和语法衔接)，汉语中指代、省略和连接手段占 59.6%，英语占 77.0%。本文的标注内容包括全部语法衔接，且包含大部分衔接手段，具有一定的代表性。

现有的对齐语料库中，仅仅有句子等单位对齐，而没有衔接的对齐，这直接影响汉英衔接对齐知识的获取。本文标注了子句、指代、省略和连接及其对齐信息。如例 2 的标注内容见图

1，图 1 中用相同颜色表示对齐的子句，用连线表示衔接对齐的信息，如连接词“尽管”和“Despite”对齐；用括号表示省略的信息，省略的内容可以是连接词，也可以是指代词，如：省略的内容“并且”和“and”对齐；同一语言中的指代链，用虚线表示，如“污染”和“污染”，“pollution”和“the problem”在分别在同一指代链上。实际标注中，指代、省略和恢复是相互指导，交叉进行的。

图 1. 例 2 的标注信息

汉英篇章衔接对齐语料库的对齐标注总原则是“单位对齐，词对齐”。标注语料的整体策略是源语为主，目标语为辅，即以汉语为主，英语为辅。标注目标是实现双语衔接中的子句、指代、连接的对齐标注。所以它实质上是一个“标注中有对齐，对齐中有标注”的对齐与标注合二为一的过程。

汉英篇章衔接的对齐标注，包括切分（子句）对齐、连接词对齐、指代对齐这几个关键对齐标注任务，由于本文考查的省略主要是连接词省略和指代省略，因此将其标注合并到相应的项目中，在标注时体现省略信息。下面详述其标注策略。

3.1 子句对齐标注

基本假设是具有对译关系的篇章，其内部的子句是一一对应的，参考李艳翠等（2013）的子句定义。英汉双语篇章子句的对齐，为了保持结构的一致性，一般采用“源语优先”即汉语优先的划分子句的方法，首先按既定的汉语基本篇章单位进行切分，然后以英语对齐（最终可根据结果归纳英语基本篇章单位）来保证汉英篇章的对应关系。根据子句定义，英语的从句或句子和子句对应，子句对齐后便于衔接信息的对齐标注。本文子句以汉语为主，将英语相应的组块（英语从句或短语）和汉语子句对应，事实上，这种分析对于汉语是子句分析，对于英语则主要是子句对齐。这种分析机制，可以保证所研究的问题是篇章层面的问题。

在实际操作中，主要依据三点：第一主要看英汉的句意。对于一个优质的翻译文本，源语中的因果、转折、并列等逻辑语义关系必然在目的语中得到反映，根据逻辑语义关系，可以分别从英汉平行语料库中相邻的子句中找出其对应关系，从而进行英汉的对齐划分；第二看结构，结合源语与目的语的结构，英汉中主谓宾的顺序是一致的，一些名词性从句、状语从句的对译也较为一致，找出英汉中相应的词汇从而找出英汉相对应的句子成分进行划分。比如看源语中结尾的动词、非谓语动词、宾语、各种从句或是其他成分在汉语中是否得到了体现；第三是看标点，在对译的中文语料库中，英文的标点大部分会和汉语一致，根据标点情况，可以更清楚地推测文意及翻译的中文文本。

如例 3，汉语子句“比开放前的一九九一年增长九成多。”和英语子句“growing more than 90 % compared to 1991 , before they had opened .”对应。

例 3a：中国十四个边境对外开放城市一九九五年经济建设取得可喜成果。|据统计，这些城市去年完成国内生产总值一百九十多亿元，|比开放前的一九九一年增长九成多。

例 3b: In 1995 , the economic construction of China 's fourteen border municipalities that are open to the outside attained gratifying results .| According to statistics ,these municipalities last year fulfilled more than 19 billion yuan of the gross domestic product ,| growing more than 90 % compared to 1991 , before they had opened .

3.2 连接词对齐标注

句子之间或子句之间存在如条件、转折、因果等语义连接关系，连接词指具有子句及其以

上语法单位连接和关系提示作用的语言单位，可以根据连接词的管辖范围(连接的子句)和篇章关系两方面确定连接词。李艳翠等（2015）将连接词作为篇章关系的关键因素在汉语中已进行了标注，参考汉语篇章结构中的做法，在汉英连接词对齐标注时，对连接词是否可添加和可删除进行标记，为便于操作，本文仅对在汉语、英语或汉英中都出现的连接词进行标注，对双语均省略的连接词，由于添加时所选择的词范围较大，容易导致对齐标注不一致，且在实际应用中意义不大。汉英对译篇章由于意义相同，所以对于连接词的汉英对齐标注主要为管辖范围和逻辑功能的对齐，标注时如连接词缺省则根据意义对连接词进行添加，对汉英都无法添加连接词的情况不进行标注。

李艳翠等（2015）在汉语连接词分类中认为，连接词可分为四大类：并列类、转折类、解说类和因果类，在此基础上又可分为 17 种不同的关系类型。例如，并列类可分为并列关系、顺承关系、递进关系、选择关系和对比关系五种关系类型。每种关系类型又包含多个连接词，而某些连接词可属于不同的关系类型。标注时主要考虑三种连接词对齐关系，如例 4 汉语没有连接词而英语有连接词，如例 5 汉英均有连接词，如例 6 汉语有而英语没有连接词。

例 4a:其中，台湾对祖国大陆输出值为一百七十八亿美元，比上一年增长百分之二十；|输入值为三十一亿美元，比上年增长百分之七十四。

例 4b:The number of investment projects dropped by 444 as compared with last year ,| **but** the value of investments rose by more than 130 million US dollars as compared with last year .

例 5a：并投资一千三百多个亿，加强基础设施和基础产业建设，|**为**扩大对外开放创造良好环境。

例 5b:It has invested more than 130 billion yuan to strengthen the construction of infrastructures and basic industries| **so as to** create a sound environment for expanding the opening up to the outside world.

例 6a:在投资项目上比上年减少四百四十四件，|**但**投资金额却比上年增加一点三亿多美元。

例 6b:Last year, the number of investment proposals presented by Taiwanese businesses and approved by Taiwan authorities totaled 490 ,| with a value of 1.092 billion US dollars.

在翻译时，允许出现不是一对一的情况，如例 7：

例 7a:在社会主义市场经济体制建设不断推进，对外开放进一步扩大的新形势下，海关的职能**不能**削弱，|**只能**加强。

例 7b:Under the new circumstances in which the construction of a socialist market economy mechanism is continually being promoted and the opening up to the outside world is further expanding, the functions of Customs **should not be** weakened, |**and should only be** strengthened .

3.3 指代对齐标注

经过反复的研究和实践，最终确定汉英篇章衔接对齐标注总原则，以篇章为单位将 ACE 实体类型为人名、地名、机构名、时间等具有代表性的且在文章中出现频率较高的指代实体词进行汉英对齐标注。标注原则是单语中的指代信息构成指代链，汉英指代链中的项目两两相互对应。标注过程中采用边标注指代链边进行双语对齐，标注和对齐同时进行，这样可以全面考察双语的各种信息。

本文标注实体指代和事件指代信息，如例 8 的“金川公司”是实体代词，“这里”“这家企业”是事件指代。例 8a 中的“金川公司”“这里”“金川公司”和“这家企业”分别对应例 8b 的“Jinchuan Company”、“this place”、“the Jinchuan Company”和“this enterprise”，同时形成指代关系，在本篇章中都指的是“金川公司”，因此将有指代信息汉英指代词标注在同一指代链。

例 8a: 一九六四年，**金川公司**产出第一批电解镍。从此以后，逐步改变了中国镍、钴及铂族金属长期依赖进口的局面。如今，**这里**已成为中国最大的镍钴生产基地和铂族金属提炼中心，镍和铂族金属产量分别占全国的百分之八十八和百分之九十以上，被誉为中国的“镍都”。一九

七八年，**金川公司**被中国政府列为全国矿产资源综合利用三大基地之一，作为中国镍工业代表的**这家企业**由此踏上依靠科技进步求振兴的发展之路。

例 8b: In 1964 , **Jinchuan Company** produced the first batch of electrolytic nickel .From then on , the situation of China 's long time dependence on import for nickel , cobalt and platinum family metals has been changed gradually .Up to now , **this place** has become China 's largest nickel and cobalt production base and platinum family metals refining center , with an output of nickel and platinum family metals that respectively account for more than 88 % and 90 % of the whole country respectively , being praised as China 's " Nickel Capitol " .In 1978 , **the Jinchuan Company** was listed by the Chinese government as one of the top three bases of integrated utilization of national mineral resources .Since then , **this enterprise** , as a representative of China 's nickel industry , began to step onto its vigorous development road by relying on advances in science and technology .

3.4 省略对齐标注

省略可以包含代词的省略、名词的省略以及连接词的省略等，本文认为指代和连接都可以省略。由于对篇章的理解是主观的，因此特别是翻译者的主观理解将会添加到翻译后的文本中，以更好的反映原文，因此省略处理的原则是汉英都省略的不做处理，主要处理汉语或者英语省略。由于汉语省略较多，标注时以英语为主，在汉语中寻找对应，如不存在则补充，存在则对齐，如不能补齐，则对空。如图 2 例 3a 中，根据英语对照补充两个省略的代词“他”(例 3a 中用“()”标示)，“(他)-he”、“他-he”、“阿 Q-Ah Q”以及“(他)-he”依次对齐。如图 1 中的例子“and”在是翻译时补充的内容，可以分析得出汉语中省略了对应的词“并且”。当然，也有一些词是汉语中有，而英语在不影响理解的情况下省略，此时英语中也补充并对齐。

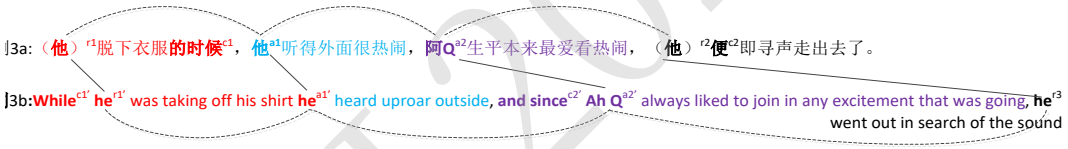

图 2. 例 1 省略和指代的对齐标注

4 汉英衔接对齐资源创建

本文将充分利用已有的汉语篇章级资源，在 OntoNotes 的汉英平行文本上追加与篇章衔接性相关的指代、省略和连接标注信息并进行汉英标注内容的对齐。为了便于标注，基于标注策略，制定了标注规范，开发了辅助标注平台，并以人工和计算机结合的方式进行语料标注。

4.1 语料选择

在 OntoNotes 中已经包含实体、部分省略信息。但这些信息是单语标注，没有体现双语对齐关系。本文在此基础上添加其他衔接信息，同时考虑双语，标注的同时完成对齐，具体包括：1）将汉英篇章中的子句标注扩展到双语；2)连接词及其对齐标注，以前期研究为指导，标注连接词属性和对齐信息，包含添加的连接词和连接词是否可删信息，连接词的管辖范围，连接词所连接的篇章单位是否调序等；2)种类齐全的汉英省略信息：OntoNotes 语料中仅包含了主语位置的零指代关系，而汉语省略涉及多个种类，这里主要标注指代和连接两种省略信息。

4.2 标注规范

根据篇章衔接分析机制和对齐策略，针对子句、连接词、指代、省略的标注及对齐分别提出具体的标注规范。标注规范注重可操作性，分别从判定原则、对齐方法等方面入手制定，并制定了标注规范。

4.3 标注方法

在标注规范的指导下进行标注，标注工作参考了之前汉语篇章结构语料资源构建积累的方

法和经验，分阶段进行：在第一阶段，由于语料库处理工作量大，为了确保质量和通用性，制定了初步的标注规范，同时开发了标注工具，并对参与标注的人员进行了培训；第二阶段，为保证标注的一致性，将标注者分为三组，分别标注了若干篇相同的文档，然后在一起讨论所有标注内容，包括指代、省略和连接的属性和对齐方式等，形成统一的标注思想，而后得到修订后的标注规范；第三阶段，标注者分组完成 60 篇相同文档的标注，用标注完的文档两两计算标注的一致性。选取一致率高的两组语料，由标注成员共同参与讨论，经过多次研究形成最终的标注规范；第四阶段，根据最终的标注规范，由标注一致率高的两组成员继续完成剩下语料的标注，另一组成员负责完成语料校对和一致性的计算，形成最终的汉英篇章衔接对齐语料库。

对于子句、指代、省略和连接及其对齐信息的标注，本文开发了辅助标注平台，根据用户选择记录需要添加的词、标注信息的类型，对齐的位置等信息，使用人机结合的标注策略，提高标注质量和效率。

4.4 标注结果

完成了 200 个平行文档的汉英篇章衔接对齐语料标注。标注了 200 个平行文档的子句切分、连接词对齐和指代词对齐语料。根据制定的汉英子句对齐切分标准，通过汉英子句对齐的标注规范，即对平行语料库进行汉英子句对齐语料标注。目前平行语料中共有效标注汉语句子 899 句，英语句子 1281 句，汉英 2153 个子句对，汉语子句平均长度是 11 个词语，英语子句平均长度是 20 个单词。汉语子句对应的英语子句主要句法结构有 S、VP、NP、PP 等。连接词对齐标注中，共标注了 817 对连接词，如“但”和“nevertheless”对应，共标注显式连接词 462 次，出现次数较多的连接词（并 and）占 50.9%，汉语中隐性连接词达到 60%。指代对齐标注中，目前共标注文档有效文档 193 篇，标注了 1613 个指代链，平均每篇文档有 8.4 个。共标注了 3657 个指代词，平均每个指代链上有 2.3 个指代词。省略情况主要是连接词省略和指代省略，在连接词省略中，中文省略 122 个词，英文省略 3 次，中文省略现象明显多于英文。指代省略 114 次，其中中文省略 92 次，英文 22 次。

5 实验结果与分析

5.1 标注质量评估

一致性评估主要考察标注者标注的一致内容与所有标注内容之比，本文从汉语一致性、英语一致性和汉英对齐一致性三方面进行考察。其中，汉英对齐一致性指的是标注者对相同语料的汉语标注一致并且汉语相对应的英语对齐标注也一致的情况。标注工作有 6 名同学参与，前期将 6 名同学两两分为 A、B 和 C 三组进行标注，对其标注的 60 篇文档进行逐一探讨并两两计算一致性，得出 A-C 小组在在汉语一致性、英语一致性和汉英对齐一致率等方面明显高于其他两个小组，因此由 A-C 小组继续完成剩下文档的标注工作，B 小组成员负责校验。由于标注内容不同，针对子句、连接词和指代词分别采用了不同的计算方法。目前共完成 200 篇文档的标注工作，其子句对齐、连接词对齐和指代对齐语料评估结果如表 1 所示。

		汉语一致性	英语一致性	汉英对齐一致性
子句对齐	切分对齐 I	0.972	0.992	
	切分对齐 II	0.968	0.930	0.909
连接词对齐	显隐对齐	0.962	0.987	0.974
	显式连接词对齐	0.800	0.950	0.876
	全部连接词对齐	0.678	0.690	0.684
指代词对齐		0.933	0.932	0.920

表 1.标注一致性计算结果

子句对齐亦可称作切分对齐，切分对齐的方法有两种：切分对齐方式 I：汉语子句的切分位置均标有标点符号，并计算了用作切分标记的标点符号（，；：。）一致性。英语子句切分不一定使用标点符号作为切分标记，可以使用空格（基本上是任意单词或标点符号）的形式作为切分标记，以及是否可以使用任何空格作为一致性计算的切分标记；切分对齐方式 II：计算不同标注者的所有切分（AUB）之间的共同切分（A∩B）的一致性。对于句子位置 SentencePosition = “ X1 ... X2 | Y1 ... Y2”，计算 A 和 B 的切分位置相同的情况。与切分对齐方式 I 相比，该方法的评估更准确，可以统一中英文切分评估标准。

从表 1 可以看出，子句切分对齐 I 在汉语和英语一致性上较高，主要是每个切分位置都进行计算，计算的无歧义切分位置较多。采用子句切分 II 计算出汉英对齐一致性为 90.9%，说明子句完全对齐还有待提高，可以从提高英语切分对齐标注的位置精准性和在汉语指导下进一步实现英语切分对齐这两方面的改进可以有效改善切分对齐标注准确率。

由于连接词总是有一定的管辖范围，且连接词有显隐之分。连接词对齐标注评估，从显隐对齐、显式连接词和对齐全部连接词对齐三个方面进行评估。由表 1 一致性结果可知，显隐对齐一致率较高，其中英语一致率达 0.987，同时英语普遍高于汉语的一致率。这是因为英语显式连接词明显较汉语的多，相比汉语，英语对于连接词有比较共性的认识，汉语的认识却有较大分歧。这也证明英语在关系对齐标注时作为指导性标准的可靠性。显式连接词对齐的一致性高于全部连接词，主要是表示同中连接关系所添加的隐式连接词不固定，如表因果可以是“因为”、“因”等词。为提高连接词对齐标注的准确率，可从两方面入手：第一，进一步明确汉语连接词的定义，从而增强汉语显式连接词的对齐标注效果。第二，规范隐式连接词的添加，指定添加连接词的范围，减少隐式连接词添加的分歧。

指代词对齐主要计算标注者选择指代词的一致性，由于指代词通常比较明显，添加的指代词多为名词且固定，所以一致性高于连接词对齐。汉英指代词对齐标注的一致性达 0.920 在指代对齐标注一致性计算中除对汉语一致、英语一致、汉英对齐一致率进行计算，还加入了汉语位置一致、英语位置一致、属性一致、指代词个数一致和指代链个数一致率的计算，其对应的一致率分别为：0.926 、0.925 、0.931 、0.932 、0.872，其一致率的计算对汉英篇章衔接对齐语料库的构建具有重要的参考意义。由于两小组同学进行双盲标注，标注结果存在一定差异。讨论过后，将进一步规范标注策略，对一些文档标注完善，个别误差大的文档进行重新标注。在对结果进行一致性评估时，将考虑去除一些无效指代链，将进一步提高一致率的精确性。

5.2 简单实验结果

李艳翠等（2013）在基于逗号的汉语子句识别研究中，手工标注了 100 篇文档。实验结果表明，具有最佳识别效果的最大熵分类器模型使用 CTB6.0 提供的标准语法树，最高准确率为 92.8%，使用 Berkeley 自动语法分析树，最高准确率是 89.9%。本文开发了汉英子句切分平台和英语子句切分平台，利用最大熵、决策树、贝叶斯等模型进行训练，然后分别进行汉语、英语子句的自动切分。得到中文自动切分准确率 90%，英文 93%。在此基础上，进行基于 BiLSTM-CRF 模型进行分析，汉英子句切分 P、R、F 分别为 92.3%、94.4%、93.4%和 95.5%、93.4%、94.4%。中文连接词自动识别准确率为 92.5%，中文 95.7%。

汉英连接词的自动识别实验，中文连接词自动识别准确率为 92.5%，中文 95.7%。李艳翠等（2015）在标注了 CTB6.0 中 500 个文档的实验结果表明，具有最佳识别效果的解说类的准确率为 82.5%，连接词自动识别并分类的总正确率为 89.1%。本文的关系识别采用英文连接词本身和对应的中文连接词作为特征，通过实验统计，给定连接词中，并列类有 332 个，占 71.86%；解说类 43 个，占 9.31%；转折类 23 个，占 4.98%；因果类 64 个，占 13.85%。通过表 2 中结果看到，并列类，因果类和解说类分类结果较好，转折类识别效果较差。

关系类别	正确率	召回率	F1 值
并列类	95.45	90.00	92.59
因果类	94.43	77.38	83.52
转折类	30.00	88.14	44.28
解说类	95.80	67.03	72.79

表 2.连接词识别结果

实验发现，由于在关系类别分布中并列类所占比例最高，训练实例最多，并且连接词的集中度较高，因此识别率相对较高。但是存在连接词一对多的现象，如“and”在并列类中出现 217 次，在解说类和因果类中各出现 1 次，所以会导致一些识别错误。转折类识别效果最差，一是因为关系类别分布中转折类出现次数最少，只有 23 次，二是因为有的转折类连接词同时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。而且观察测试结果发现也将该“but”归为了转折类。解说类虽然比例较低，在训练集中共出现 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。根据实验分析，与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，连接词类别识别准确率就越好。以后通过增加训练集规模，训练结果会得到大幅的提升。

6 标注中的难点问题及解决

6.1 标注对象问题

在最初的标注过程中，发现标注结果中真正形成指代链的实体词较少，并且存在较多指代词单独成链的现象，最终造成不同标注者的标注结果存在较大差异。经过反复的实践和讨论，最终统一标注规范，将有较多指代词的 ACE Type 为 GPE、ORG、LOC、PERSON 和 DATE 的实体词标注处理，存在较少实体词甚至往往仅有单独一个实体词的 ACE Type 为 MONEY、PERCENT、EVENT、QUANTITY 和 CARDINAT 等实体词不再单独标注成链。

例 9a: （中国）h1 羽绒及其制品行业是（八十年代中期）d1 开始快速发展的，全行业利用（中国）h2 资源、人力优势，加上注重引进国外先进技术与设备，产品产量和质量得以大幅度提高。据不完全统计，目前（中国）h3 已有羽绒及制品加工企业（三千余家）c1，其中上规模的达（六百多家）c2，从业人员约（三十万）c3，形成年产羽绒制品（五千多万件）c4 生产能力，年工业总产值达（八十亿元）c5。通过（十余年）d2 市场开拓，（中国）h4 现已成为世界主要羽绒生产国和羽绒制品出口国，年出口羽绒近（三万吨）c6、羽绒制品（二千多万件）c7，创汇达（八点二亿美元）c8，其中羽绒服装出口额占行业出口总额（百分之五十）c9 以上。

例 9b:**China 's**[h1] down and down products industry started its rapid development **in the mid '80s**[d1] .The entire industry makes use of **China 's**[h2] resources and manpower advantage , and additionally stresses introducing advanced foreign technology and equipment , thus increasing production volume and quality by a large margin .According to incomplete statistics , **China**[h3] currently has over **3,000**[c1] down and down product enterprises , among which , those above scale have reached more than **600**[c2] , with employed staff of about **300,000**[c3] .It has an annual production capacity of **50 million**[c4] down products with a total annual industrial output value reaching **8 billion yuan**[c5] .Through more than **ten years '**[d2] market development , **China**[h4] has now become the world 's main down manufacturing country and down products export country , annually exporting nearly **30,000 tons**[c6] of down and over **20 million**[c7] down products , with earned foreign exchange reaching

820 million US dollars[c8] , including down clothing export values accounting for more than **50 %**[c9] of total industry export values .

例 9 中 ACE Type 为 GPE 的实体词有（h1~h4）,依据对齐标注原则，该实体词可标注成指代链。其中 ACE Type 为 DATE 的实体词有 d1 和 d2,因其仅有一个实体词，不单独标注成链。ACE Type 为 CARDINAT 的实体词(c1~c4 和 c7) 、ACE Type 为 MONEY 的实体词 c5 和 c8、ACE Type 为 QUANTITY 的实体词 c6 以及 ACE Type 为 PERCENT 的实体词 c9 不在要求标注的实体词范围内，同样不单独标注成链。

6.2 特殊语境指代词标注难点

例 10a: 近年来，(中)e1(韩)q1 两国之间的经贸往来发展迅速。截止去年九月，(韩国)q2 在(华)e2 投资企业总数为五千八百八十三家，(中国)e3 已成为(韩国)q3 最大的投资对象国。据(中国)e4 海关统计，一九九五年两国贸易额已达一百六十九点八亿美元，比前年增长百分之四十四点八。经济专家预计，今年(中)e5(韩)q4 两国贸易额将增至二百五十亿美元。

例 10b:In recent years , the economy and trade contacts between the countries of **China**[e1] and **South Korea**[q1] have been developing rapidly .By September of last year , the total number of **Korean**[q2] enterprises investing in **China**[e2] totaled 5,883 .**China**[e3] has become **Korea 's**[q3] largest target country for investment .According to **Chinese**[e4] Customs statistics , in 1995 , trade between the two countries reached 16.98 billion US dollars , increasing 44.8 % compared with that of the previous year .Economic experts estimate that this year trade between the two countries of **China**[e5] and **South Korea**[q4] would increase to 25 billion US dollars .

例 10 中的 “中”（e1）、“华”（e2）和“中”（e5）若单独出现时，并不能准确判断其具体含义。在本篇文章中，根据其在文章的语境，以及上下文信息,很容易判断其与“中国”（e3 和 e4）形成指代衔接，将其（e1~e5）标注在同一指代链，对应的英文中正确翻译出“china”。同样“韩”（q1）和“韩”（q4）与“韩国”（q2 和 q3）形成指代衔接，应将其（q1~q4）标注在同一指代链，对应英文翻译“South Korea”。

7 结语

本文进行了汉英篇章衔接语料库的标注工作，主要实现了子句、连接词、指代和省略的对齐标注。汉英篇章衔接对齐语料库的对齐标注总原则是“单位对齐，词对齐”，标注语料的整体策略是以汉语为主，英语为辅，省略添加的原则是汉语或英语有对应显式词出现。子句以汉语为主，将英语相应的组块（英语从句或短语）和汉语子句对应。连接词对齐标注连接词及其语义关系，根据其体现为管辖范围和逻辑功能的对齐。单语中的指代信息构成指代链，汉英指代链中的项目两两相互对应，汉英都省略的不做处理，主要处理汉语或者英语省略。在本文汉英衔接对齐标注策略基础上，选择汉英平行文本进行了汉英篇章衔接资源的构建，目前完成了 200 篇平行文档的标注工作。标注中采用辅助平台，对子句、连接词、指代的标注质量分别进行评估，评估结果说明本文方法切实可行，简单实验结果表明本语料子句切分、连接词识别具有较强的可计算性。下一步工作将不断完善本标注策略，扩大标注语料，进行指代和省略的计算分析工作。

致谢

本文得到国家自然科学基金项目（61502149）、河南科技学院高层次人才科研项目（2017039）、华东师范大学统计与数据科学前沿理论及应用教育部重点实验室开放课题资助。实验室小组吕天赐、李书磊、李强、胡大帅、侯昆昊和王基翱等同学认真负责的参与了标注工作。感谢公开学习资源的专家学者，感谢提出建设性指导意见的 CCL 2020 的专家学者。

参考文献

Cook G. 1989. *Discourse*. Oxford Univ Pr (Sd).

Halliday，M．A．K．＆R．Hasan．1976. *Cohesion in English*．Edward Arnold , London.

Halliday，M．A．K．1994. *An Introduction to Functional Grammar*．Edward Arnold , London.

Huang H H, Chen H H．2011.*Chinese discourse relation recognition Proceedings of the 5th International Joint Conference on Natural Language Processing*． Asian Federation of Natural Language Processing． Chiang Mai. 1442-1446.

Kong F. and Zhou G. D. 2010.*A Tree Kernel-based Unified Framework for Chinese Zero Anaphora Resolution.* In Proceedings of EMNLP 2010.

Li Y C，Feng W H，Kong F，et a1．2014.*Building Chinese discourse corpus with connective-driven dependency tree structure.* In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP)．Doha．2105-2114.

Li Y，Feng W，Sun J，et a1．2014.*Building Chinese dis— course corpus with connective-driven dependency tree structure.* In Proceedings of EMNLP 2014. 2105-2114.

Louis A，Joshi A，Nenkova A．2010.*Discourse indicators for content selection in summarization Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*．Portland：Association for Computational Linguistics．147-156.

McCallum A K．2002.*Mallet：a machine learning for language toolkit*．(2002)[2012-02-28]．http：// mallet．CS．umass．edu.

Pradhan S., Hovy E. and Marcus M. et al. 2007. *OntoNotes: A Unified Relational Semantic Representation.* Iternational Journal of Semantic Couputing, 1(4):405-419.

Pradhan S., Ramshaw L., and Weischedel R. 2007. *Unrestricted Coreference: Identifying Entities Events in OntoNotes*. In Proceedings of ICSC'2007.

Tu M., Zhou Y. and Zong C. Q. 2014. *Enhancing Grammatical Cohesion: Generating Transitional Expressions for SMT.* In Proceedings of ACL'2014.

Werth P. 1984.*Focus, Coherence and Emphasis*. Routledge Kegan & Paul.

曹继阳.2019.*汉语口语语篇衔接手段与衔接成分——基于经典情景喜剧《我爱我家》的研究*.语言文字应用,(2):142.

常阳.2015.*英语倒装结构在汉英篇章翻译中的衔接功能及应用*.北方文学（中旬刊）,(12):70-71.

冯洪玉,李艳翠,冯文贺.2019.*基于汉英平行语料库的英文显式篇章关系识别*.河南科技学院学报（自然科学版）47(5):55-62.

冯文贺,李艳翠,任函,等.2017.*汉英篇章结构平行语料库的对齐标注评估*.中文信息学报,31(3):86-93.

胡壮麟.1994.*语篇的衔接与连贯*.上海外语教育出版社,上海.

孔芳,王红玲,周国栋.2019.*汉语篇章理解研究综述*.软件学报,30(7):2052-2072.

李艳翠. 2015.*汉语篇章结构表示体系及资源构建研究*.苏州大学,江苏.

李艳翠,冯文贺,周国栋,等.2013.*基于逗号的汉语子句识别研究*.北京大学学报（自然科学版）, 49(1):7-14.

李艳翠,孙静,冯文贺,等. 2015.*基于连接依存树的汉语篇章结构分析平台*.中国中文信息学会.中国中文信息学会 2015 学术年会(CIPS2015)暨第十四届全国计算语言学学术会议(CCL2015)、第三届基于自然标注大数据的自然语言处理国际学术研讨会(NLP-NABD2015)论文集.1-10.

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

王菲.2018.*从衔接论看汉英语篇翻译中的衔接与连贯——以《落花生》为例*.青春岁月,(11):136-137.

奚雪峰,孙庆英,周国栋.2019.*面向意图性的篇章话题结构分析研究与展望*.计算机学报, 42(12):2769-2794. DOI:10.11897/SP.J.1016.2019.02769.

徐凡,朱巧明,周国栋,等.2014.*衔接性驱动的篇章一致性建模研究*.中文信息学报,28(3):11-21,27.

杨传鸣.2008.*《红楼梦》及其英译本语篇衔接对比*.黑龙江大学,哈尔滨.

杨凤丽.2012.*论汉英否定篇章衔接功能的对比*.齐齐哈尔大学学报：哲学社会科学版,(2):155-156.

张献丽.2017.*略论汉英翻译中的衔接性*.牡丹江大学学报,26(10):146-147,150.

张易男,李燕鸿.2019.*汉英"照应"衔接对比与翻译研究——以《2018 年政府工作报告》及其英译版为例*.英语教师,19(9):134-138,141.

钟书能.2016.*话题链在汉英篇章翻译中的统摄作用*.外语教学理论与实践,(1):85-91,58.

周利芳.2018.*汉语"提及"类衔接成分的用法及其辨析*.华文教学与研究,(3):61-69.

朱永生,郑立信,苗兴伟.2001.*英汉语篇衔接手段对比研究*.上海外语教育出版社,上海.