一个面向信息抽取的中英文平行语料库

惠浩添[1,2]，李云建[1,2]，钱龙华[1,2]，周国栋[1,2]

（1.苏州大学 自然语言处理实验室，江苏 苏州 215006；
2.苏州大学计算机科学与技术学院，江苏 苏州 215006）

摘要：除了机器翻译，平行语料库对信息检索、信息抽取及知识获取等研究领域具有重要的作用，但是传统的平行语料库只是在句子级对齐，因而对跨语言自然语言处理研究的作用有限。鉴于此，本文以 OntoNotes 中英文平行语料库为基础，通过自动抽取、自动映射加人工标注相结合的方法，构建了一个面向信息抽取的高质量中英文平行语料库。该语料库不仅包含中英文实体及其相互关系，而且实现了中英文在实体和关系级别上的对齐。因此，该语料库将有助于中英文信息抽取的对比研究，揭示不同语言在语义表达上的差异，也为跨语言信息抽取的研究提供了一个有价值的平台。

关键字：命名实体；语义关系；双语映射；平行语料库

中图分类号：TP391　　**文献标识码：**A

A Chinese-English Parallel Corpus for Information Extraction

HUI Haotian[1,2], LI Yunjian[1,2], QIAN Longhua[1,2], ZHOU Guodong[1,2]

(1.Natural Language Processing Lab of Soochow University, Suzhou, Jiangsu 215006, China;
2.School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: In addition to machine translation, parallel corpora play an important role in information retrieval, information extraction and knowledge acquisition etc. However, traditional parallel corpora are aligned at sentence level, thus their significance for research on cross-language natural language processing is limited. In view of this, this paper, on the basis of the OntoNotes, constructs a high quality Chinese and English parallel corpus for information extraction by combining automatic extraction, automatic mapping and manual annotation. The corpus contains the entities and their mutual relations, and achieves the alignment between Chinese and English on entity and relation level. Therefore, this corpus will facilitate comparative study of information extraction in Chinese and English, reveal the difference of semantic expression between languages, and also provide a valuable platform for research on cross-language relation extraction .

Key Words: Named Entity; Semantic Relation; Bilingual Mapping; Parallel Corpus

1 引言

信息抽取是指从自然语言文本中抽取出有用的实体、关系和事件等信息，并把它们存放到一个结构化的数据库中。根据 ACE 的定义[1]，信息抽取包括三个主要任务：命名实体识别（Named Entity Recognition）、实体关系抽取（Relation Extraction）和事件抽取（Event Extraction）等。信息抽取对问题回答、文本摘要、信息融合、知识获取等自然语言处理应用领域有着重要的研究意义。

收稿日期：2015-05-31　　**定稿日期：**2015-08-07
基金项目：国家自然科学基金(61373096，90920004)；江苏省高校自然科学研究重大项目(11KJA520003)

主流的信息抽取研究都采用统计机器学习方法，因而语料库的规模和质量对信息抽取的性能至关重要，但是人工标注大规模的语料库是一件费时又费力的事情。另一方面，在自然语言处理中往往存在着多种语言的可比较或平行语料库，有效利用这些多语言语料库是提高信息抽取性能的途径之一。Chen 等[2]在中英文平行语料之间进行命名实体的联合识别和对齐，旨在同时提高两种语言的命名实体识别性能。Kim 等[3]利用平行语料库来实现从英文到韩文的跨语言关系抽取，即将源语言中识别出来的实体和关系映射到目标语言中。Qian 等[4]利用机器翻译的方法将中英文语料库互相翻译，并将一种语言的实体和关系映射到另一种语言中，从而同时促进两种语言中关系抽取的性能。

上述研究说明，平行语料库对于提高跨语言信息抽取的性能具有重要的作用，但是目前的平行语料库一般都在句子级对齐[5-7]，并没有实现在实体和关系级对齐，因而在实体和关系的双语映射过程中存在着一定的错误。而在信息抽取中广泛使用的 ACE 语料，尽管标注了多语种的实体和关系，但并不是平行的。为了弥补现有平行语料库中存在的不足，本文从 OntoNotes 中英文平行语料库出发，以 ACE 2005 中文语料库的标注规则为基本指南，通过自动抽取和手工标注相结合的方法构建了一个面向信息抽取的平行语料库。尽管受 OntoNotes 语料库的限制，该平行语料库的规模比较小，我们仍希望该语料库可以为揭示中英文语言表达上的差异和跨语言信息抽取的研究提供一个基准的平台。

2 中英文平行语料库的构建

本节首先对标注目标和任务进行说明，然后指出标注过程中的挑战以及解决办法，接着详细描述中文语料库的构建方法，最后评估语料库的一致性。

2.1 标注目标和任务

本文的目标是要构建一个面向信息抽取（主要是实体和关系）的中英文平行语料库，以便于中英文双语实体识别和关系抽取的研究。该语料库应包含完整的中英文对齐的实体和关系标注信息[1]。实体标注信息包括实体类型、指称范围、指称级别和实体类别等；关系标注信息包括关系类型、句法结构、关系时态等。除此之外，还应该实现指代链的标注。下面是一对中英文平行句对：

（c1）[乍得]$_{1\text{-}1}$ 新 [总统]$_{2\text{-}2}$ [依迪斯·代比]$_{2\text{-}3}$ 十二日 到达 [巴黎]$_{3\text{-}4}$ 访问 。[密特朗]$_{4\text{-}5}$ [总统]$_{4\text{-}6}$ 同 [他]$_{2\text{-}7}$ 进行 了 半 小时 秘密 会谈 。

（e1）[Chad]$_{1\text{-}1}$'s New [President]$_{2\text{-}2}$ [Idriss Deby]$_{2\text{-}3}$ arrived in [Paris]$_{3\text{-}4}$ on the 12th for a visit . [President]$_{4\text{-}5}$ [Mitterrand]$_{4\text{-}6}$ had a half - hour 's secret meeting with [him]$_{2\text{-}7}$.

其中，方括号内的内容表示实体指称，下标表示其编号，而底划线表示左面实体和右面实体具有一定的语义关系。该平行句对中包含 7 个实体指称，4 个实体（其中 2-2，2-3 和 2-7 是实体 2 的不同指称，4-5 和 4-6 是实体 4 的不同指称），2 个语义关系（实体 1 和实体 2 之间具有 ORG-AFF.Employment 关系；而实体 2 和实体 3 之间具有 PHYS.Located 关系）。

2.2 标注难点及解决方法

一般而言，只要找到中英文平行语料库，然后参考 ACE 的标注规范分别进行实体及其关系的标注，最后将实体和关系对齐即可得到中英文平行语料库。但这样做需要大量的人力，花费的时间也很长。本文从下面四个方面来讨论在标注中遇到的关键问题及其解决方法。

2.2.1 语料库的选择

传统的面向机器翻译的平行语料库虽然数量很多，但均没有在双语上对齐的实体及关系标注信息，将它们标注成面向信息抽取的双语平行语料库工作量太大。本文选择 OntoNotes 中的

[1] 目前还没有考虑事件标注信息。

新华社中英文平行语料（共有 325 篇文章），它不仅具有较高的句子对齐率，而且也标注了部分实体信息，这将显著减轻标注工作量。不过，即使是这样，通过对已有标注信息的观察，我们发现还存在着以下的问题：

1) 实体指称类型单一：OntoNotes 仅标注了命名实体，即指称级别为 NAM 的实体，所有的名词性指称(Nominal)和代词指称(Pronoun)均没有标注，这不符合一个面向信息抽取的语料库的要求；
2) 指代链未完全合并：虽然 OntoNotes 标注了指代链信息，但不完整。比如句子“据 [泰国] 官员 透露 ， １９９５年 ， 缅 [泰] 两 国 贸易 总额 超过 3 亿 美元 。”中的“泰”和“泰国”应属于同一个实体，但目前的标注并没有合并到同一个指代链中。

为了解决上述问题，同时减轻标注工作量，本文遵循“自动+手工”的原则来构建双语平行语料库，充分利用 OntoNotes 语料的平行句对和现有标注信息，其主要步骤包括两个方面：

1) 中文语料的标注：即从中文 OntoNotes 语料库中产生已标注的实体信息，调整中文实体标注信息，标注中文实体间语义关系；
2) 英文语料的映射：即将中文的实体及其关系标注信息映射到英文中，并调整英文的实体及关系标注信息。

2.2.2 实体的嵌套

实体嵌套是一个比较普遍的现象，比如中文短语“[[宁波]国际发展信托投资公司]”中包含了两个实体。在 ACE 的标注规范中，为了简化问题将它作为一个实体，即不考虑被嵌套的实体。这样做的缺点是丢失了许多命名实体及其语义关系，因为嵌套的实体之间一般都存在的语义关系，这将会对今后的命名实体识别及关系抽取任务造成一定的影响。本文考虑了中文的实体左嵌套现象和英文的左右嵌套现象，从而提高了语料中命名实体和实体关系的数量，同时也便于今后命名实体识别及关系抽取工作的进行。

2.2.3 实体类型的辨析

某些实体在不同的上下文中会呈现不同的角色，例如 GPE 类型的实体可以代表相应的地区、组织或人物，在 ACE 的标注规范中以角色来表明这种差别。我们发现另一个实体类型 ORG 也具有相似的特点，例如：

今天 在 [上海 国际 金融 学院] 正式 举行 开学 典礼 ，参加 开学 典礼 的 有 [学院] [院长] ...

其中，实体“上海国际金融学院”在前一子句中强调设施，因此具有 FAC 的角色，而在后一子句中则表示 ORG 本身。ACE 标注规范没有区分这种差异，从而在某些情况下导致 ORG 和 FAC 类型出现混乱。为了解决这个问题，本文对 ORG 实体类型同样引入了角色这个概念，它包含 ORG 和 FAC 两种角色。

2.2.4 关系类型的辨析

在某些情况下，区别实体关系类型变得很困难。在 ACE2005 的中文语料库中，不同的标注者对类似的语言表达式往往给出不同的语义类型，甚至同一个标注者也会出现不一致的情况。为了提高标注的一致性，本文整理了易混淆的关系类型对，并针对它们分别制定了可操作的标注规则。表 1 列出了这些易混淆的关系类型对及其区分规则。

表 1 易混淆的关系类型及其区分规则

序号	关系类型 1	关系类型 2	论元类型	区分规则
1	ORG-AFF.Membership	ORG-AFF.Employment	PER 与 ORG	若为会员制且不支付工资，则为关系 1,否则为关系 2。
2	PART-WHOLE.Geographic	PHYS.Near	GPE 与 LOC	一个国家或地区的边界、边境为关系 1，两个国家或地区的边界、边境为关系 2。

3	PART-WHOLE.Subsidiary	GEN-AFF.Org_Location	ORG 与 GPE	GPE 的管理、行政、执法等机构为关系 1，否则关系 2。
4	PHYS.Located	ART.UOIM[2]	PER 与 FAC	人物单纯位于某个场所，而非使用其特定功能为关系 1，否则为关系 2。

2.3 中文语料的标注

首先从 OntoNotes 中抽取中文实体标注信息，但由于这些标注信息极不完整，因此还需要手工调整实体标注信息，并标注实体间语义关系。

2.3.1 中文实体标注信息的产生

中文 OntoNotes 语料以嵌入标记的形式标注了文本中出现的命名实体和指代链，分别以后缀名 name 和 coref 存储在两个文件中，其中指代链中不仅标明了实体的指代关系，也标明了概念之间的指代关系。标注信息的产生过程包括以下三个步骤：

1) 从 name 文件中读出实体标注信息；
2) 从 coref 文件中读出实体的指代链信息；
3) 将实体标注信息和实体指代信息合并为统一的标注信息。

2.3.2 中文实体及关系的标注

为了方便快捷地标注实体和关系信息，我们利用 Java 语言开发了专门的标注工具，其主界面如图 1 所示。它由左右两个部分组成。左面是所有文件列表，右面是该文件所对应的文本内容，其中不同的前景色表示不同类型的实体，而两个实体之间的底划线表示它们之间存在语义关系。

当要增加和修改实体指称，进入图 2（a）所示的实体标注界面，标注者可以调整实体类型、实体类别和指称级别等信息。要增加和修改实体关系时，进入图 2（b）所示的关系标注界面，标注者同样可以修改关系的类型、句法结构和时态等信息。

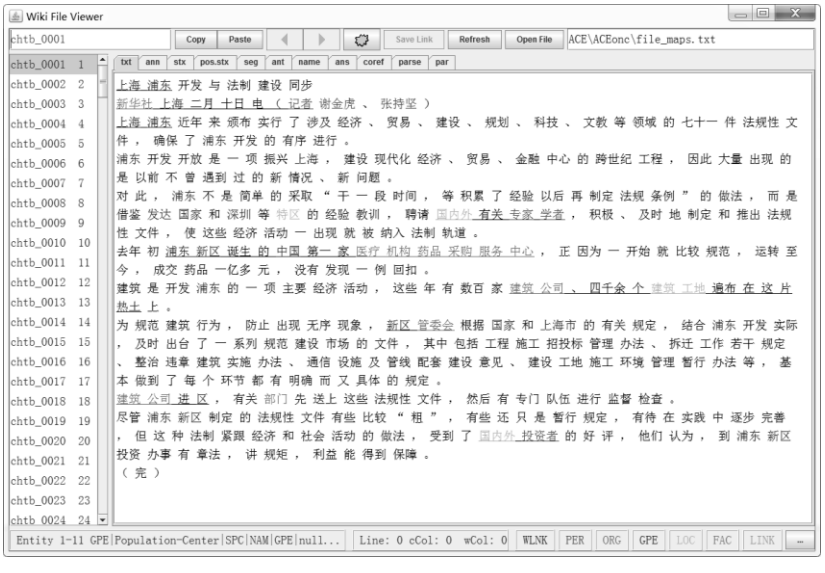


图 1 主界面

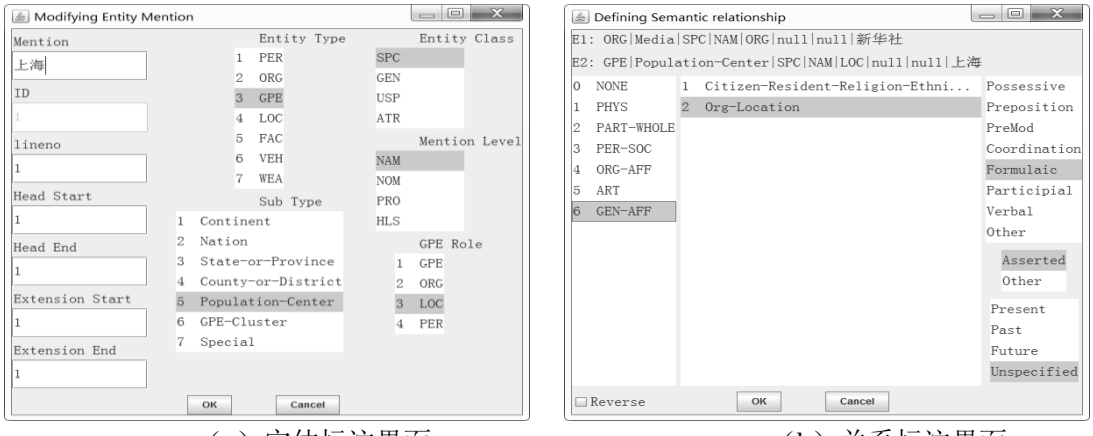
2.4 英文语料的映射

[2] 全称为 User-Owner-Inventor-Manufacturer

在中文实体和关系标注完之后，就需要把这些信息映射到英文中，从而获得英文的实体及关系标注信息。这个映射分为三个步骤：句子对齐、实体映射和关系映射等。

2.4.1 句子对齐

OntoNotes 语料库虽然是中英文平行的，但它只提供了文件之间的对齐，并没有提供句子之间的对齐关系，因此映射的第一步便是句子对齐工作。考虑到新华社新闻语料的翻译质量较高，本文采用相似度计算方法来实现自动句子对齐，即逐个比较中英文句子之间的相似度，然后再人工调整对齐结果。



（a）实体标注界面　　（b）关系标注界面

图 2 实体和关系标注界面

1. 句对相似度的计算

在计算中英文句子之间的总体相似度时，考虑了表 2 所示的四种相似度，并对它们进行加权平均，即：

$$S_T = \sum_{i=1}^{4} w_i * S_i$$

其中 S_T 为总体相似度，S_i 为某一个特征的相似度，w_i 为该相似度的权值，权值由实验来确定。

表 2 句子对齐的相似度特征

特征符号	特征名称	特征描述	权重
S1	句子位置	句对在文本中位置的最小值与最大值之比	0.2
S2	句子长度	句对中词次个数的最小值与最大值之比	0.1
S3	所含实体数	句对中实体个数的最小值与最大值之比	0.1
S4	所含词次数	句对中各个词次两两相似度累加，再进行归一化	0.6

2. 平行句对的产生

得到中英文句子间的两两相似度后，就可以在此基础上进行句子对齐。平行句子的对齐方法有动态编程及分裂聚类策略[8]、基于词汇的 Champollion 对齐方法[9-10]和针对非单调句子的半监督对齐方法[11]。考虑到语料库的质量较好，中英文之间严格按照句子顺序对齐，因此本文采用分段对齐法，其基本思想是首先找出相似度最高的句对作为平行句对，然后用该句对分隔句子范围，再在各自范围内继续匹配。具体算法如图 3 所示。

算法：句子对齐

输入： Sim[M][N]，中英文句子间的两两相似度，M为中文句子数，N为英文句子数

输出： Pairs[]，中英文平行句对集合

初始化： Ranges([1, M]，[1, N])，中英文句子范围

步骤：

从Ranges中弹出句子范围range：

在范围range中根据找出相似度最高的句对(i，j)；

将(i，j)加入到句对集合Pairs中；

将range按照(i，j)分隔成上下两个了范围，各自加入到Ranges中；

直到为空

图 3 句子对齐算法

需要说明的是，为了避免相似度过低的句对被识别为平行句对，本文设置了根据实验获得的最低阈值 α=0.35，低于该阈值的不能作为平行句对。

3. 人工调整

由于采用 OntoNotes 新华社新闻专线中的 325 篇平行语料，翻译质量较高，因而句子对齐率较高（约 95%），人工调整并不需要消耗太多的时间和精力；而且其英文翻译语法规范、句法结构清晰，这将非常有利于实体和关系的映射。

2.4.2 实体映射

在实体对齐之前，首先要进行词对齐。常用的词对齐的算法有 Brown 等[12]提出的 IBM 模型和 Vogel 等[13]提出的隐马尔科夫模型。另外，Feng 等[14]提出了最大熵结合自举算法进行命名实体对齐。本文是将自动抽取及人工标注的中文实体映射到英文中，因而先采用 Giza++工具进行词对齐，然后再映射实体及人工调整。考虑到 OntoNotes 的平行语料库规模不大，可能会影响到词对齐效果，因此本文将 OntoNotes 语料和 FBIS 语料结合起来一起进行词对齐，最后再将其分离开单独处理。对于词对齐的效果，本文从中随机抽取 25 句对进行分析，这里以中文为源语言、英语为目标语言，最终的词对齐准确率约 80%，召回率约 72%，造成召回率较低的原因主要是中英文语言的差异。

词对齐完成后，接着便是实体对齐。由于并非所有的实体指称都是单个词次构成，所以本文利用如下启发式规则：

1) 中文实体指称的词次连续，则对应英文实体指称的词次也必将连续；
2) 不存在多个中文实体指称对应一个英文实体指称。

根据以上两个启发式规则，将中文中的实体尽可能地映射到英文中，就初步得到英文语料中的实体标注信息。为了对实体对齐的正确率进行分析，本文随机抽取 13 篇进行分析，发现实体对齐的准确率约 79%，召回率约 73%，这与词对齐的效果相差无几。这说明基本上是词对齐错误导致了实体的丢失。因此，下一步的工作是人工进行进一步调整，最终实体对齐率可以达到 93%左右。

2.4.3 实体关系映射

实体映射及其手工调整完成之后，接下来的关系映射就比较简单，但也需要考虑以下三个问题：

1) 关系实例的两个论元必须处于一个句子中。由于平行句对中存在一对多的情况，原来中文中处于同一句的两个实体有可能映射到两句不同的英文中。在这种情况下，丢弃该关系实例；

2) 关系实例的两个论元的前后顺序是否交换。如果交换了顺序，则必须改变关系类型的正逆性；

3) 关系映射后的句法结构是否变换。由于中英文对同一语义关系的表达方式存在差异，因此关系实例的句法结构可能会发生变化，并且也无法准确预测新的句法结构，因此对关系实例的句法结构有必要进行人工调整。

至此，经过中文语料标注和英文语料的映射后，包括实体和关系对齐信息的中英文平行语料库就全部构建完毕。

2.5 语料标注的一致性

语料标注的一致性体现了标注的难度和语料的质量。为了保证标注质量，我们招募了两名志愿者，分两个阶段标注中文实体及其关系：

1) 第一阶段：两名志愿者首先对 25 篇文章中的实体或关系进行标注，然后由一名仲裁者检查标注的差异，改正共同的错误，并允许存在有争议的差异，最后计算两名标注者之间的一致性；

2) 第二阶段：两名志愿者分别标注剩下的 300 篇文章，每人大约标注一半。

在衡量实体标注的一致性时，只考虑实体指称的中心词和实体大类，采用常规的准确率（P）、召回率（R）和调和平均（F1）；在衡量关系标注的一致性时，只考虑关系小类，同样采用常规的准确率（P）、召回率（R）和调和平均（F1）。表 3 列出了两名标注者在调整前后的实体和关系标注的一致性指标。

从表中可以看出，调整前实体的召回率较低，这是因为两位标注者对实体标注的某些要求（如实体类别等）不够了解。而经过调整后，无论是实体还是关系的一致性已达到可接受水平。

表 3 实体和关系标注的一致性

一致性项目	P(%)	R(%)	F1
实体（调整前）	88.9	79.8	84.1
实体（调整后）	98.8	98.4	98.6
关系（调整前）	82.3	73.5	77.6
关系（调整后）	95.5	92.9	94.2

3 平行语料库统计分析

为了更好地揭示中文和英文在表达实体、关系等方面的语言差异，本文分别就对齐率、实体指称缺失情况、关系句法结构的变化等三个方面进行统计和分析。

3.1 中英文对齐率

为了考察在中文到英文的对齐过程中标注信息的保留情况，表 4 统计了实体指称、实体和关系在中文中的数量，对齐到英文后的数量以及对齐的百分比。由于在英文的翻译过程中，很多文本标题行被省略了，从而导致平行句对的丢失，因此为了分析标注信息丢失的真正原因，表中也列出了在句子对齐情况下的统计数据。例如，“全部实体指称”是指语料库中标注的所有实体指称，而下面一行“全部实体指称（句子对齐）”表示出现在平行句对中的实体指称。从表中可以看出：

表 4 实体指称与实体总数统计

对齐类别	中文	对齐（英文）	对齐率(%)
全部实体指称	15,784	14,738	93.4
全部实体指称（句子对齐）	14,982	14,738	98.4

全部实体	6,963	6,865	98.6
全部实体（句子对齐）	6,918	6,865	99.2
全部关系	5,147	4,726	91.8
全部关系（句子对齐）	4,899	4,726	96.5

1） 实体对齐率最高，实体指称对齐率次之，而关系对齐率最低。这是因为只要实体的任一个指称能对齐，则实体就能对齐；而只有一个关系的两个实体指称都对齐，关系实例才能对齐。

2） 无论对于何种统计指标，句子对齐情况下的对齐率均高于全部语料库情况下的对齐率，并且对齐率均超过 95%。这说明如果仅考虑平行句对中的对齐情况，那么可以认为标注信息的对齐是相当成功的。因此，在后续表格中，本文丢弃非平行句对中的标注信息，从而便于更准确地分析语言之间的真正差别。

3.2 实体指称缺失

从表 4 中可以看到，在实体指称映射中存在缺失现象，即一个中文实体指称没有对应的英文实体指称，从而影响到关系的对齐。表 5 把 244 个实体指称的缺失原因进行分类，并列出了各个原因所占的比例。

表 5 实体指称缺失原因

缺失类型	实体数目	百分比（%）
语义	140	57.4
句法	32	13.0
翻译	68	27.9
规则	4	1.6
合计	244	100.0

由表 5 可以发现，约 2/3 的实体缺失是由于中英文语言差异造成的，而约 1/3 的实体缺失是由句法和翻译问题所致，只有极少部分为规则不允许中间嵌套造成的，具体为：

1. 语义缺失

语义缺失是指缺失的实体被本句中的其他指称表述，并不需要再赘述；或者是本句中的某个实体可以暗含多个实体。例如，在（c2）句中，[中国]与[自己]为“中国”的不同指称，而在（e2）句中，“中国”一词的指称，并未像中文句子中出现两次，这是因为在英文中一个指称完全可以表达句意。

（c2）[中国] 愿意 为 不断 加深 这 种 友谊 作出 [自己] 的 努力 。

（e2）[China] is willing to make efforts to continually deepen this type of friendship .

2.句法缺失

句法缺失是指由于中英文在词法和句法上的差异而导致的实体指称的丢失，约占到 10%以上，分析表明，其原因有两个方面：一是专有名词缩写，即中英文在某些专有名词缩写上具有一定的差异性，即某些中文名称是从英文缩写中翻译过来的。例如，中文“ [联合国] [安理会]”中包含两个实体，而其对应的英文“ [UNSC]” 却只有一个实体。二是 HLS 表述差异，所谓 HLS 引用类多集中在 “...个，...的，...家，...之一”等词，而在英文中并未有与“个”，“的”，“家”相对应的词。

3. 翻译缺失

有将近三分之一的实体缺失是由于翻译原因而引起的，即英文中未将相应的中文实体翻译出来，而且并不能被其他实体的指称所表述或暗含。例如，在（c3）中的[河南省] 并未在（e3）中出现。

（c3）记者 从 [[河南省] 文物 考古 研究所] ...其中 有 肋骨 、 趾骨 等 。

（e3）This reporter has learnt from the [Archaeological Institute of Cultural Relics] ... such places as Hutou Hill , Yangcheng Township , Xixia county , etc .

4. 规则问题

为了尽可能多地标注嵌套实体，同时也便于处理，我们规定对于中文实体只考虑左嵌套情况，而对于英文实体，左嵌套和右嵌套都要考虑，这就导致某些实体无法对齐。例如在“[[上海]施贵宝]”中存在两个实体，而在其英文“[Squibb 's -LRB- Shanghai -RRB-] ”中，由于“Shanghai”这个实体没有出现在最右侧，因此不被标注为一个实体。

3.3 关系句法结构

分析中英文在实体关系语言表达方式上的句法差异对关系抽取研究具有很好的指导作用。表 6 统计了中文关系实例映射到英文关系实例时句法结构发生变化的实例数量，其中行和列分别表示中英文句法结构类型，需要注意的是英文比中文多出两个句法结构类型，即所有格和介词。从表中可以看出：

表 6 中文到英文关系句法结构的转换

句法结构	所有格	介词	前修饰	并列	公式	分词	谓词	其它	小计
前修饰	342	1240	1370	6	–	20	9	96	3083
并列	–	–	–	2	–	–	–	1	3
公式	–	–	–	–	738	–	–	–	738
分词	1	63	1	–	–	53	1	28	147
谓词	2	36	–	–	–	36	322	130	526
其它	1	41	1	–	–	5	4	177	229
小计	346	1380	1372	8	738	114	336	432	4726

1) 多于一半的中文前修饰结构发生了变换，主要变换为介词（约 40%）以及所有格结构（约 11%），并且当前者发生时，往往还伴随着关系论元先后顺序的交换。例如，中文中的“[外交部] [副部长]”，英文翻译为[vice minister] of the [Ministry of Foreign Affairs]，句法结构由前修饰转换为介词。
2) 中文的公式结构映射到英文时，仍然为公式结构。这是由于公式结构通常都是新闻报道中的固定模式，即使翻译成英文，也不会发生变化。
3) 相当一部分中文分词结构（超过 40%）转换为英文的介词结构。这是由于中文中前置的分词结构（如“驻”、“在”、“来自”和“遍布”等）在英文中往往被翻译成后置的介词结构或分词结构，因此两个关系论元的位置也会发生变化。
4) 在中文谓词结构中，也有少部分转换为英文的介词结构（约 7%）和分词结构（约 7%）。例如，在（c4）-（e4）中，由谓词结构转换为介词结构。而在（c5）-（e5）中，由谓词结构转换为分词结构。

（c4）最后 一 批 俄罗斯 [军队] 撤离 [德国] 的 仪式 3 1 日 在 柏林 举行 。

（e4）*The ceremony for the withdrawal of the last group of Russian [troops] from [Germany] was held in Berlin on the 31st .*

（c5）*[德国] 领土 上 存在 [占领军] 的 状态 行将 结束 。*

（e5）*The [occupying armies] existing in [German] territory will end soon .*

最后一个值得注意的现象是，由于中文句法结构到英文句法结构的转换在各个类型并不均匀，因而导致中英文关系实例中句法结构的主导类型不同。在中文中，约 65%的关系实例都是前修饰结构；而在英文中，前修饰结构和介词结构的关系实例均占 29%左右。不同的句法结构可能会导致中英文关系抽取的难度不一样。

4 总结与展望

本文在已有的 OntoNotes 中英文平行语料库基础上，结合 ACE 实体和关系标注中存在的问题，制定了一些额外的标注原则，通过自动抽取和映射，再加人工调整的方法完成了一个包含实体和关系对齐信息的中英文平行语料库，该语料库具有较高的标注一致性。通过对语料库的统计表明，尽管从中文到英文的翻译过程中存在着成分缺失的现象，但实体对齐率和关系对齐率均达到了 95%以上，这说明平行句对之间的关系信息能基本保留；另一方面，中英文语言在表达语义关系的句法结构上有一定差异，中文有 65%以上都通过前修饰结构来表达，而英文则还通过介词结构来表达。

今后的工作，我们将利用本文构建的实体关系平行语料库，比较中英文关系抽取的差异性；还将利用该平行语料库进行跨语言信息抽取等方面的研究，如双语协同训练、双语主动学习等。

参考文献

[1] Doddington G R, Mitchell A, Przybocki M A, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation[C]//LREC. 2004:837-840.

[2] Chen Y, Zong C, Su K Y. On jointly recognizing and aligning bilingual named entities[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 631-639.

[3] Kim S, Jeong M, Lee J, et al. Cross-Lingual Annotation Projection for Weakly-Supervised Relation Extraction[J]. ACM Transactions on Asian Language Information Processing (TALIP), 2014, 13(1): 3:1-3:26.

[4] Qian L, Hui H, Hu Y, Zhou G, Zhu Q. Bilingual Active Learning for Relation Classification via Pseudo Parallel Corpora[C]//ACL 2014:582-592.

[5] Xiao R. The Babel English-Chinese parallel corpus[DB/OL]. [2013-02-13]. http://www.lancaster.ac.uk/fass/projects/corpus/babel/babel.htm.

[6] Ma, Xiaoyi. Hong Kong Parallel Text LDC2004T08[DB]. Web Download. Philadelphia: Linguistic Data Consortium, 2004.

[7] United States. Joint Publications Research Service, United States. Foreign Broadcast Information Service. JPRS Report: China[M]. Foreign Broadcast Information Service, 1993.

[8] Deng Y, Kumar S, Byrne W. Segmentation and alignment of parallel text for statistical machine translation[J]. Natural Language Engineering, 2007, 13(03): 235-260.

[9] Ma X. Champollion: A robust parallel text sentence aligner[C]//LREC 2006: Fifth International Conference on Language Resources and Evaluation. 2006: 489-492.

[10] Li P, Sun M, Xue P. Fast-Champollion: a fast and robust sentence alignment algorithm[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 710-718.

[11] Quan X, Kit C, Song Y. Non-Monotonic Sentence Alignment via Semisupervised Learning[C]//ACL (1). 2013: 622-630.

[12] Brown P F, Pietra V J D, Pietra S A D, et al. The mathematics of statistical machine translation: Parameter estimation[J]. Computational linguistics, 1993, 19(2): 263-311.

[13] Vogel S, Ney H, Tillmann C. HMM-based word alignment in statistical translation[C]//Proceedings of the 16th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1996: 836-841.

[14] Feng D, L ü Y, Zhou M. A New Approach for English-Chinese Named Entity Alignment[C]//EMNLP. 2004, 2004: 372-379.

作者简介：

惠浩添（1991——），男，硕士研究生，主要研究领域为信息抽取。
Email: 20134227019@stu.suda.edu.cn；（通讯作者）

李云建（1991——），男，硕士研究生，主要研究领域为信息抽取。
Email: 20145227020@stu.suda.edu.cn；

钱龙华（1966——），男，副教授，硕士生导师，主要研究领域为自然语言处理。
Email: qianlonghua@suda.edu.cn；

周国栋（1967——），男，教授，博士生导师，主要研究方向为自然语言处理。
Email: gdzhou@suda.edu.cn。