语篇标注语料库的建设研究[1]

练睿婷　史晓东

厦门大学人工智能研究所，福建厦门，361005

E-mail：lianlian1022@gmail.com　mandel@xmu.edu.cn

摘要： 本文主要介绍一个语篇标注体系，该标注体系采用自底向上的方法对基本语篇形式单元（BFDU）到句群等语篇中不同层次的语言单位和其间的关系类型进行标注，并标注了能充分反应语篇信息的各种词汇关系。本研究所产生的语料库可用于基于语篇的机器翻译、句法分析、信息抽取等多种应用领域的训练语料和测试语料。

关键词： 语篇标注，语篇基本形式单元(BFDU)，语料库

The Construction of Discourse Annotation Corpus

Lian Ruiting, Shi Xiaodong

Institute of Artificial Intelligence, Xiamen University, Xiamen Fujian 361005, China

E-mail：lianlian1022@gmail.com　mandel@xmu.edu.cn

Abstract: This paper presents a discourse annotation system. The system is based on bottom-up approach. It annotates the language unit and the different types of relationships in different levels from basic forms of discourse unit (BFDU) to sentence groups. It also annotates various kinds of word relationships which can reflect the discourse information. The annotated corpus can be used as the training and/or testing corpus in a variety of applications such as discourse-based machine translation, parsing and information abstracting.

Keywords: Discourse Annotation，Basic Formal Discourse Unit(BFDU)，Corpus

引言

在自然语言处理领域，语料库是构建语言数学模型的基础。许多成功的机器翻译系统的核心模型也都是在语料库的基础上建立起来的。但是语篇标注的语料库还处于起步阶段，其标注的对象和方法等都有待于研究。目前在语篇标注中，很多是以信息抽取、问答系统、文本摘要为目标应用，在机器翻译方面研究不多[1]。

本文介绍的语篇标注体系是针对基于语篇的机器翻译提出的一个适合英汉双语的标注体系，是基于语篇的机器翻译关键技术研究的一项开端性工作，但该研究所产生的语料库也不仅限于机器翻译，也可用于其他应用系统如文本摘要、文本信息提取等。

本文的结构如下：第 1 节简要介绍国内外的相关工作以及给我们研究的启发。第 2 节按照自底向上顺序来介绍我们的语篇标注体系所要标注的对象以及标注的方法。第 3 节列出了标记集及相关的实例。最后小结，并提出下一步的工作和目标。

1. 研究背景

[1]本文受到国家自然科学基金（No. 60573189），国家 863 计划（No. 2006AA01Z139），国家 863 计划（No. 2006AA010108-3），省重点项目（No. 2006H0038），省基金项目（No. 2006J0043）资助。

06 年 ACL 会议在语篇标注方面有一个 Tutorial 称为 Discourse Annotation: Discourse Connectives and Discourse Relations[2]。相应的语篇标注模型有宾州语篇标注树库 Penn Discourse TreeBank (PDTB)[3]，其思想是将连词（Connectives）看成是二元语篇关系的谓词，通过标注连词的论元 Arg1 和 Arg2 来标注语篇关系。其他的语篇标注模型还有基于 Rhetorical Structure Theory 的 RST Treebank[4]，基于 Linguistic Discourse Model (LDM) 的标注分析[5]，基于 Segmented Discourse Representation Theory 的标注分析[6]等。以上的模型都试图把语篇划分成一些表示意义的基本单元，标注这些单元之间的句法语义关系，构建语篇的某种意义表示，从而达到对语篇的全局了解[1]。只是这些在语篇标注的基本单元上定义有所不同，PDTB，RST 等认为子句是基本语篇单元（Elementary Discourse Unit，EDU），但是不包含做主语、宾语的嵌入式子句等。LDM 中则认为 EDU 是一个语义单位，这些语义单位呈现一定的句法结构，不一定是子句，也可以是一个时间状语，一个连词，而且 EDU 之间可以嵌套，其划分要比 PDTB 和 RST 细。受这些启发，我们的标注体系在语篇标注的基本单元上也做了相关的研究。

语篇的关系不仅仅只有子句或者 EDU 之间的关系，语篇中的各种词汇关系也是语篇理解的一个重要手段。相关的研究有 Automatic Content Extraction (ACE)[7]系列评测，其标注的内容包括语篇中的各种实体、事件、关系、时间、数值等，还包括一些指代现象。WordNet[8]和 HowNet[9]虽然并不是在语篇层次上来描述词和和概念的语义，但是其中定义的一些词汇关系都可以在同一个语篇中体现，如果在语篇中标注这些相关的词汇关系，则更能体现关于词或词组的句法信息。受这些启发，我们的标注体系在语篇词汇关系上也做了相关的研究。

目前尚未知将词汇关系、子句关系、句子关系等各种可利用关系共同标注的语篇标注的相关工作，事实上对于语篇的理解都需要利用这些关系，因此，建设这样的语篇标注语料库对于基于语篇的机器翻译、文本摘要、信息抽取等应用系统是很有必要的。

2 语篇标注体系介绍

我们的语篇标注是一个从基本语篇形式单元到句群标注的自底向上的标注过程，我们的语篇标注所要标注的内容有：

1）语篇基本形式单元及其类型的标注

2）子句的标注以及子句间关系的标注

3）句子的划分以及句子间关系的标注

4）句群以及句群中心句的标注

5）语篇词汇关系的标注

下面就从这几个方面来简要介绍我们的语篇标注体系。

语篇基本形式单元及其类型的标注

由于汉语中的“一逗到底”现象比较多，我们的标注体系中针对汉语的这个现象提出了语篇基本形式单元的概念，并对其类型进行标注。这样的标注可以为长难句的句法分析等服务。

2.1.1 语篇基本形式单元的定义和类型

我们认为，以标点符号（顿号，引号等除外）隔开的分句或句子片断是语篇的基本单位，称为语篇基本形式单元 Basic Formal Discourse Unit(BFDU)。BFDU 和语义无关，语篇的分析在这

个基础上进行比较直观。其划分只需简单的以标点符号划分，这些标点不包括冒号、顿号、引号、括号以及在引号和括号内的句子标点。

由于 BFDU 的粒度差别很大，可能是短语、子句、句子，也可能是句法结构不完整的片段，我们根据 BFDU 的句法结构是否完整将 BFDU 划分为合法 BFDU 和不合法 BFDU。这样的标注有助于长句的句法分析。

我们的语篇标注体系暂定在宾州树库[10, 11]上进行标注，判断合法 BFDU 和不合法 BFDU 可以借助树库来进行，如果树库中的该片段能构成一棵子树，则判断为合法的，否则判断为不合法的。我们采用宾州树库的语法体系，形式定义如下：

合法 BFDU（well-formed BFDU）定义为：如果某 BFDU 在正确的句法树中能构成一棵子树，而不是森林结构，则该 BFDU 为合法 BFDU。这样的 BFDU 直观的可理解为子句、短语。

不合法 BFDU（ill-formed BFDU）定义为：如果某 BFDU 在正确的句法树中不能构成一棵子树，则为不合法 BFDU。

2.1.2 语篇基本形式单元的合并

不合法的 BFDU 无论对于句法分析还是对后面句群的划分和对分析句子和句子之间的关系都没什么意义，所以我们在标注的过程中对其进行合并，直到合并到是一个合法的新 BFDU 为止。

2.2 子句的标注以及子句间关系的标注

对机器翻译而言，子句是一个重要的概念。翻译的时候，子句之间基本上存在着对应关系。因此，我们的标注体系中除了标注 BFDU 外还明确的标注子句以及子句间的关系。

2.2.1 子句的定义

子句（clause）是一组包含一个主词和一个动词的关连字，一个子句一定是包含一个动词，即使该动词是省略掉的。如：【I will work harder】 if 【I can】. 该例句仍然有 2 个子句，第 2 个子句的动词被省略。

在我们的标注体系中，BFDU 和子句没有直接的关系。一个子句可能跨越多个 BFDU，而一个 BFDU 又可能含有多个子句。如："【He said 【he will come】】" 这是一个 BFDU，但是它含有两个动词 "said" 和 "come"，所以这个 BFDU 包含两个子句。又如："【学生们比较喜欢年轻，美丽的教师】" 这个句子含有两个 BFDU，但是只含有一个动词 "喜欢"，所以只含有一个子句。

为了更准确的划分子句，我们采用宾州树库的语法体系，即将我们这里定义的子句与宾州树库中的 IP 对应。即如果某个片段在树库中能构成一棵以 IP 为根结点的子树则为子句。

2.2.2 子句间的关系

美国语言学家 Halliday 把子句复合体与传统语法的句子相提并论，且认为子句复合体中各子句的关系划分为相互依赖型和逻辑语义关系[12]。相互依赖性关系我们定义为嵌套关系，逻辑语义关系我们又分为并列关系和偏正关系。逻辑语义关系这样划分与 PDTB[3]中将连词分成两类 Subordinating conjunctions 和 Coordinating conjunctions 是兼容的，这样我们的标注体系和 PDTB 也具有较好的可比性。子句间关系类型描述如下：

嵌套关系：通常指一些从句现象。如：定语从句、宾语从句、状语从句等。

并列关系：包括子句之间的并列、承接、连贯、递进和选择关系。

偏正关系：包括子句之间的因果、假设、条件、让步关系等非并列关系。

2.2.3 子句和子句间关系的标注

由于 BFDU 和子句的交叉现象（如 2.2.1 中的例子），我们在标注的时候将 BFDU 的标注和子句的标注放在同一层次上。这样就很好的解决了交叉的现象。我们定义的子句中是不包括连词和功能词。标注子句时，只要标注出子句在相应句子中的起始位置和结束位置即可。（参见下一节的实例）

子句间关系的标注我们将发生关系的两个子句看成是两个关系的两个参数，采用的标注格式是“关系名称（Arg-1，Arg-2）”，其中对于嵌套关系，Arg-1 定义为嵌套子句的子句（即相对大一点的子句），Arg-2 定义为被嵌套在 Arg-1 里的子句；对于并列关系， Arg-1 为相对位置在前的子句，Arg-2 就相对位置在后的子句；对于偏正关系，Arg-1 定义为正句，Arg-2 定义为偏句。

2.3 句子的划分和句子间关系标注

汉语的句子和英语的句子在形态上略有差异，汉语缺少形态变化，句子没有语法上的形式标记。为了使我们的标注体系适合汉语也适合英语，有必要对句子的划分和句子间关系进行折中的定义。

2.3.1 句子的划分

句子是语言运用的基本单位，它由词或词组构成，能表达一个完整的意思。我们的标注体系中句子的划分采取了既适合汉语又适合英语的划分方法，即凡是以句末标点（句号、问号、感叹号、省略号）结尾的片段，都划分为一个句子。

2.3.2 句子间关系

句子间关系是在连贯的语篇中句子与句子之间在结构上和意思上的联系。不同的语言学家对句子间关系的类型定义都有所不同，我们这里将句子间关系类型分为并列关系和偏正关系。这样定义既简洁又有利于自动识别，而且还和子句间的关系保持一致性。

句子间关系类型描述如下：

并列关系：包括句子之间的并列、承接、连贯、递进和选择关系。

偏正关系：包括句子之间的因果、假设、条件、让步关系等非并列关系。

2.3.3 句间关系的标注

句间关系的标注的格式类似子句间关系的标注(2.2.3)。采用格式“关系名称(Arg-1,Arg-2)”，对于并列关系，Arg-1 为相对位置在前的句子，Arg-2 就相对位置在后的句子；对于偏正关系，Arg-1 定义为正句，Arg-2 定义为偏句。

2.4 句群以及句群中心句的标注

句群作为一组语义上连贯的句子的概念是值得重视的。句群提供了更多的上下文，可望在词义消歧，指代生成，时态计算方面为机器翻译提供新的思路。[13]因此，有必要对句群进行标注，从而可对句群的结构做跨语言对比研究，为语篇机器翻译及其他应用服务。

2.4.1 句群的切分

句群是在语义上有逻辑关系、在语法上有密切联系、在结构上有衔接连贯的一群句子的组合[13]。董振东在[14]中指出，句群“指的是一个完整的段落或者一个段落内若干个连贯的句子”，还指出“包含 6~8 个句子”的句群对机器翻译比较理想。黄曾阳在[15]中认为句群是“围绕着一个特定概念展开的话语”，还指出在一个句群之内的句子存在着省略和照应等形式现象，并且不同语言的形式标志不同，“英语偏好照应，汉语偏好共享”。

我们认为，句群内的句子间一定存在逻辑关系（并列或偏正）和一些词汇上的交叉等形式现象，否则，如果某两个相邻的句子不存在这些现象，就在这两个句子间作为划分句群的标记。为了方便机器的识别和标注的一致性，我们还定义了“最小化原则”即尽量把句群划分到不能再小为止，同时规定句群内句子的个数不能超过 7 句。

2.4.2 句群中心句的确定

标注句群的中心句也是很有价值的，如可以用于做摘要，也有利于对文章主题的把握。句群的中心句是表达句群中心的句子。对于由一些并列关系的句子组成的句群，我们认为都是中心句。句群的中心句可通过一些规律和标志来确定，我们在我们的标注规范中有详细的说明。

2.5 语篇词汇关系的标注

基于语篇的词汇关系，如相关词汇对与词义消歧的作用、单词重复或照应对定指名词短语（用什么冠词）的影响、单复数的判定，对机器翻译而言相当重要。对汉语（和日语）而言，省略的处理也十分关键，直接关系到翻译质量的好坏。[1]

2.5.1 语篇词汇关系类型的确定

语篇词汇关系非常复杂，其关系类型也没有明确的分类。著名语言学家 Halliday 和 Hasan 在[16]中分为 5 种：照应、省略、替代、连接和词汇衔接。Michael Hoey 在[17]中又提到：“在各种衔接手段中，词汇衔接占 48%， 照应占 36%，省略占 12%，替代占 4%”。其中这里的照应的定义就是我们通常所说的广义的共指。

我们的标注体系中的语篇标注体系主要考虑跨句的词汇关系，所要标注的词汇关系有：省略、共指、词汇衔接、相关词，其具体描述如下：

省略（Ellipsis）：我们只标注主语省略的情况，不包括共享主语的情况。

共指（Reference）：指现实世界同一实体不同描述的一组词之间的关系。共指包括：名词重复、用代词代替、用简称代替。

词汇衔接（Cohesion）：是指通过词汇选择在篇章中建立一个贯穿篇章的链条。[16]词汇衔接关系包括：反义(Antonym)、同义或近义(Synonym / near-synonym)、上下位(Hyponymy)、整体与部分(Meronymy)、转喻(Metonymy)。

相关词（Correlation）：主要包括实体属性关系、属性值关系和事件角色关系。

在我们的标注规范中对这些关系做出了具体定义和详细说明。

2.5.2 语篇词汇关系的标注

对于省略的标注，我们标注时将所省略的词补上，并且标注出省略的位置。对于其他几种词汇关系的标注，我们标注时将发生关系的所有词都作为同一个关系的不同 mention 进行标注。词汇衔接的标注，还要加上子类型的标注（参见下一节的实例）

3 标注实例

我们的语篇标注目前是对宾州树库的文本（汉语采用分过词的文本）进行标注，采用 XML 语言格式，这样可以结构化、嵌套的方式描述信息内容和各种有用关系。限于篇幅，下面给出一个简单的标注实例。

例：海关 统计 表明 ， “ 八五 ” 期间 （ 一九九0年 — 一九九五年 ） ， 中国 外商

投资 企业 的 进出口 呈 直线 上升 之 势， 出口 平均 增长 百分之四十三点二， 进口 年均 增长 百分之三十八点六 。去年 实现 进出口 总值 达 一千零九十八点二亿 美元， 占 全国 进出口 总值 的 比重 由 上 年 的 百分之三十七 提高 到 百分之三十九 。

其标注结果如下表所示：

```
<P ID="P1">
- <SG ID="SG1" CENTER="S1">
  - <S ID="S1">
    - <BFDUS>
      - <NEWBFDU ID="N1" UNITED="B1,B2,B3,B4,B5">
          <BFDU ID="B1" TYPE="ILL-FORMED">海关 统计 表明 ， </BFDU>
          <BFDU ID="B2" TYPE="WELL-FORMED">“ 八五 ” 期间 （ 一九九 0年 — 一九九五年 ） ， </BFDU>
          <BFDU ID="B3" TYPE="WELL-FORMED">中国 外商 投资 企业 的 进出口 呈 直线 上升 之 势 ， </BFDU>
          <BFDU ID="B4" TYPE="WELL-FORMED">出口 年均 增长 百分之四十三点二 ， </BFDU>
          <BFDU ID="B5" TYPE="WELL-FORMED">进口 年均 增长 百分之三十八点六 。 </BFDU>
        </NEWBFDU>
      </BFDUS>
      <CLAUSES>
        <CLAUSE id="C1" start="1" end="26" />
        <CLAUSE id="C2" start="15" end="26" />
        <CLAUSE id="C3" start="27" end="31" />
        <CLAUSE id="C4" start="32" end="36" />
      </CLAUSES>
      <C_RELATION ID="C_R1" ARG1="C1" ARG2="C2" TYPE="嵌套关系" />
      <C_RELATION ID="C_R2" ARG1="C1" ARG2="C2" TYPE="偏正关系" />
      <C_RELATION ID="C_R2" ARG1="C1" ARG2="C3" TYPE="偏正关系" />
      <C_RELATION ID="C_R3" ARG1="C2" ARG2="C3" TYPE="并列关系" />
    </S>
    <S ID="S2">
      <BFDUS>
        <BFDU ID="B6" TYPE="WELL-FORMED">去年 实现 进出口 总值 达 一千零九十八点二亿 美元 ， </BFDU>
        <BFDU ID="B7" TYPE="WELL-FORMED">占 全 国 进出口 总值 的 比重 由 上 年 的 百分之三十七 提高 到 百分之三十九 。 </BFDU>
      </BFDUS>
    - <CLAUSES>
        <CLAUSE id="C5" start="1" end="8" />
        <CLAUSE id="C6" start="3" end="8" />
        <CLAUSE id="C7" start="9" end="24" />
        <CLAUSE id="C8" start="15" end="24" />
      </CLAUSES>
      <C_RELATION ID="C_R4" ARG1="C5" ARG2="C6" TYPE="嵌套关系" />
      <C_RELATION ID="C_R5" ARG1="C5" ARG2="C7" TYPE="偏正关系" />
      <C_RELATION ID="C_R6" ARG1="C7" ARG2="C8" TYPE="嵌套关系" />
    </S>
    <S_RELATION ID="S_R1" ARG1="S1" ARG2="S2" TYPE="并列关系" />
  </SG>
</P>

- <RELATION>
  - <WORD_RELATION ID="W_R1" TYPE="Cohesion" subtype="near-synonym">
      <mention LOC="B4" START="3" END="4">增长</mention>
      <mention LOC="B5" START="9" END="10">上升</mention>
    </WORD_RELATION>
  - <WORD_RELATION ID="W_R2" TYPE="Cohesion" subtype="Hyponymy">
      <mention LOC="B3" START="6" END="7">进出口</mention>
      <mention LOC="B4" START="1" END="2">出口</mention>
    </WORD_RELATION>
  - <WORD_RELATION ID="W_R3" TYPE="Cohesion" subtype="Hyponymy">
      <mention LOC="B3" START="6" END="7">进出口</mention>
      <mention LOC="B5" START="1" END="2">进口</mention>
    </WORD_RELATION>
  - <WORD_RELATION ID="W_R4" TYPE="Cohesion" subtype="Antonym">
      <mention LOC="B4" START="1" END="2">出口</mention>
      <mention LOC="B5" START="1" END="2">进口</mention>
    </WORD_RELATION>
  - <WORD_RELATION ID="W_R5" TYPE="Reference">
      <mention LOC="B3" START="6" END="7">进出口</mention>
      <mention LOC="B6" START="3" END="4">进出口</mention>
      <mention LOC="B7" START="4" END="5">进出口</mention>
    </WORD_RELATION>
  - <WORD_RELATION ID="W_R6" TYPE="Correlation">
      <mention LOC="B7" START="7" END="8">比重</mention>
      <mention LOC="B7" START="12" END="13">百分之三十七</mention>
      <mention LOC="B7" START="15" END="16">百分之三十九</mention>
    </WORD_RELATION>
  - <WORD_RELATION ID="W_R7" TYPE="Ellipsis">
      <mention LOC="B6" START="1">中国 外商 投资 企业</mention>
      <mention LOC="B7" START="1">进出口 总值</mention>
    </WORD_RELATION>
</RELATION>
```

图 1 语篇标注语料库实例

4、结语

本文主要介绍一个语篇标注体系的相关标注思路和规范，该标注体系吸收了已有的一些标注体系的多种优点，同时尽量避免了其它标注体系的缺陷。语料库加工的工作量很大，为了保证语料库加工的质量，目前我们正在开发一个辅助的标注工具，尽量减少一些由人的主观因素而造成的错误。今后我们将继续改进和完善该标注体系，尤其是丰富的语篇词汇关系，标注更多的语料，

使其早日用于机器翻译等应用系统中。

参考文献：

[1] 史晓东，陈毅东. 基于语篇的机器翻译前瞻. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议. 2006

[2] Aravind Joshi, Rashmi Prasad and Bonnie Webber. Tutorial at the Association for Computational Linguistics, Sydney, Australia. 2006

[3] Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber. The Penn Discourse Treebank. Proceedings of the Language Resources and Evaluation Conference, Lisbon, Portugal. 2004

[4] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Current Directions in Discourse and Dialogue, Jan van Kuppevelt and Ronnie Smith eds., Kluwer Academic Publishers. 2003

[5] Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. A Rule Based Approach to Discourse Parsing. In Proceedings of SIGDIAL'04. Boston, MA. 2004

[6] Jason Baldridge and Alex Lascarides. Annotating Discourse Structures for Robust Semantic Interpretation. Proceedings of the Sixth International Workshop on Computational Semantics IWCS-6, Tilburg. 2005

[7] The ACE 2007 (ACE07) Evaluation Plan. National Institute of Standards and Technology, 2007. (Available from http://www.nist.gov/speech/tests/ace/)

[8] WordNet: http://wordnet.princeton.edu/

[9] HowNet: http://www.keenage.com/

[10] Penn Chinese TreeBank: http://www.cis.upenn.edu/~chinese/

[11] The Penn TreeBank Project: http://www.cis.upenn.edu/~treebank/

[12] Halliday. " An Introduction to Functional grammar." Edward Arnold. 1985

[13] 吴为章、田小琳. 《汉语句群》，商务印书馆，2000

[14] 萧振东. 机器翻译研究的展望，《计算机世界报》，1998

[15] 黄曾杨等. 句群标注与分析，《HNC理论与实践》，第二期. 2005

[16] Halliday, Hasan. Cohesion in English[M]. London:Longman. 1976

[17] Michael Hoey. Patterns of Lexis in Text[M]. P. imprenta : Oxford. 1991

[18] 冯志伟. 从汉英机器翻译看汉语自动句法语义分析的特点和难点. 汉语计算与计量研讨会，香港城市大学. 1998

[19] 李西新. 英汉句子成分省略对比. 焦作大学学报. 2006