# 中文事件事实性信息语料库的构建方法[1]

曹媛，朱巧明，李培峰
（苏州大学 计算机科学和技术学院，江苏 苏州 215006）

**摘要：**事件事实性表达事件是否是事实的确定性程度，在文档中表现这一属性的是特定的句子结构和词汇。本文在充分研究影响中文事件事实性的句子成分的基础上，提出了五类事件事实性相关信息并给出了具体的标注规则。最后，在 ACE 2005 中文语料库的基础上完成了 Movement 事件的事实性标注，并对标注完成的语料库进行了相关的统计和分析，为后续研究提供基础。

**关键词：**事实性；语料库；标注
**中图分类号：**TP391　　**文献标识码：**A

# The Construction of Chinese Event Factuality Corpus

CAO Yuan, ZHU Qiao Ming, LI Pei Feng
(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

**Abstract:** The factuality of an event is the degree of certainty to which an event is a factual one. In context, what expresses this attribute are the specific sentence structure and vocabularies. In this paper, we make the full study of the factors which influence Chinese event factuality, then present five kinds of factual related information of events and their annotation rules. Finally, we accomplish the annotation of the Movement event in the ACE 2005 Chinese Corpus and analyze the results, which is the foundational work of many information extraction applications.

**Key words:** Factuality; Corpus; Annotation

## 1 引言

事件事实性表达了其是否是事实的确定性程度。事件的事实性是一种语义信息，对于很多自然语言处理应用来讲，都可以作为其基础支撑或者用来提高其性能。在文本理解（Text Understanding）中，事件的事实性是作为理解篇章中事件的一类非常重要的依据。因为从事实性事件推断出的结果与从可能事件推断出的结果显然有所不同。例如，在意见分析时，同样的情况，对于不同的意见持有者，它可以是事实、可能发生的以及非事实。

目前，事件事实性的研究还处于萌芽阶段，只有英文有了小规模的确定性信息语料库。Factbank[1,2,3]以 TimeBank 为基础进行标注，把事件事实性分为五个类别，包括 208 个文档，9500 个事件和超过 77000 个词。BioScope[4,5,6]标注了生物论文中的模糊限制语及其作用范围，共有近 2 万句和超过 38 万个词。GENIA[7,8]标记了事件的极性和确定性，包括 1000 多个摘要。其他还有一些相关工作，这些工作或多或少不太完整。有的只是引入了事实性相关的某

[1] **基金项目：**国家自然科学基金资助项目（61070123，61272260）；江苏省自然基金（BK2011282）；江苏省高校自然科学重大基础研究项目（11KJA520003）。
**作者简介：** 曹媛(1989-)，女，通信作者，硕士研究生，主要研究方向为自然语言处理。

些概念或者其工作与事实性有关系，并没有系统地介绍分析事件事实性或者其研究重点并不是事实性。如 Rubin [9]关注“确定（Certainty）”这个概念，并在新闻文章上做了标注实验。如 Saurí[10]等为事件标注 Modality 属性，这里的 Modality 表示事件源对其陈述事件所持的确定程度或者态度。

事件事实性的研究是一项相当具有挑战性的工作[1,2,3]。首先，事件事实性通过句子语法表达，是一种主观性信息，并不是具体的语言学系统。它的取值是一个从事实到非事实的连续空间。确定需要标注的事实性相关信息时，既要兼顾语言学的分析也要考虑一般性常识推理。其次，事件的事实性是各种信息综合作用的结果，如事件的极性、事件发生的时间、事件源对该事件的确定性程度，甚至不同的句子结构也会对事件的事实性产生影响。

本文选取 ACE（Automatic Content Extraction） 2005 中文语料库为基础语料库，标注其中 Movement 事件的事实性，并给出了标注规则和语料库的统计分析。本文选取 ACE 2005 中文语料库作为基础语料库的主要原因有：1）ACE 语料库中，有关事件和时间相关内容已经标注，不用重复标注，节省人力物力；2）ACE 语料库中语料来源广，包括了新闻报道、广播、网络日志等；3）文本数量充足，语料库规模可扩充，为将来进行半自动甚至自动标注提供支持。

本文组织如下：第二节说明事实性标注相关内容，包括事实性相关信息和标注规则的介绍；第三节介绍语料标注过程，包括标注工具的使用和标注结果的保存；第四节对已构建好的语料库进行数据统计和分析；最后第五节对现有工作进行了总结并提出了未来工作的展望。

## 2 事实性信息标注

### 2.1 事件的事实性（Event Factuality）

事件的事实性，又称确定性和真实性，它表述了事件是否是事实的不同程度。在本文中，事件的事实性不是真实世界中的事实性，是根据事件上下文，该事件对于某一事件源的事实性。例如 E1 中，“校长辞职事件”对于事件源“郑微微”是一个确定发生事件。又如 E2，“王十离婚事件”对于该报道的发表者而言是一个可能发生事件。参考英文事件事实性标注的类别[1,4,7]，本文将事件事实性分为五大类：当然发生(Fact)，当然不发生(Counterfact)，可能发生(Probable)，可能不发生(Not probable)以及不确定是否发生(Uncertain)。

*E1：郑微微说王校长辞职了。*

*E2: 王十可能和他的妻子离婚了。*

### 2.2 事实性相关信息

事实性相关信息是具体标注的对事件事实性有影响的内容，可从词汇和句子结构两个层面分析。

首先从词汇角度来看，有一类词通常以事件触发词的上层谓词的形式出现，代表事件叙述者对该事件的立场与态度。如下例 E3 中的“怀疑”，E5 中的“计划”等。这类词在本文中称之为事件选择谓词（Event Selecting Predicate），谓词有级别属性，表示不同的谓词对事件事实性的确定程度，有“确定”、“可能”、“不确定”三种取值。此外，还有一类文章级谓词，其作用范围为整篇文档，通常出现在新闻稿中，如“中央台综合报道”中的“报道”一词就是一个文章级的谓词。

事件源与事件选择谓词联系紧密，事件源是事件的叙述者，对事件态度的持有者。一个事件可以有多个事件源，本文定义了三个层次的来源：1）发布该事件的媒体/网站等（媒体源）；2)文章作者（作者源）；3）文中该事件的描述者（直接源）。下文我们主要介绍直接源的标注，如 E3 和 E 5 中的“警方”和“他”都是直接源。

此外，有一类词，它们一般是修饰触发词的副词或者助词，在语义上表示事实性的确定程度或者事件发生时间。在例 E3 中，“可能”一词表示事件“逃离”的确定性为可能。本文称这类词为程度词。程度词有时态和级别两个属性分别表示事件事实性的时态和确定性程度。时态有“过去”、“现在”、“将来”和“无”四种取值，级别有“确定”、“可能”、“不确定”和“无”四种取值。

最后一类词，它们对事件的级性有着决定性的影响，那就是否定词。比较 E3 和 E4，E4 中因为有了否定词“没有”而使其事实性从可能发生变成可能不发生。

*E3：警方**怀疑**嫌犯**可能**逃离香港。*

*E4：警方**怀疑**嫌犯**可能**还**没有**逃离香港。*

*E5：他明天**计划**去北京。*

除了词汇，一些特定的句子结构也影响事件的事实性。在本文中，提出两种对事件事实性有影响的从句类型：条件从句和目的从句。如下例 E6 和 E7 所示。

*E6：如果你来，我就走。*

*E7：为了和小三结婚，王十要和妻子离婚。*

因此，在本文中，事实性相关信息包括从词汇和句子结构两个层面，具体有：事件选择谓词、事件源、程度词、否定词和从句五类。下表 1 是对事件事实性相关信息的总结和归纳。

表 2-1 事件事实性相关信息

| 事实性相关信息 | 定义 | 属性 |
|---|---|---|
| 事件选择谓词 | 通常以事件触发词的上层谓词的形式出现，代表事件叙述者对该事件的立场与态度 | 级别：确定、可能、不确定 |
| 事件源 | 事件的叙述者，对事件态度的持有者 | 无 |
| 程度词 | 副词或者助词，表示事实性的确定程度或者事件发生时间 | 时态：过去、现在、将来、无<br>级别：确定、可能、不确定、无 |
| 否定词 | 表否定意义，用于否定一个命题 | 无 |
| 从句 | 在本文中，特指对中文事件事实性有影响的从句，即：条件从句和目的从句 | 无 |

## 2.3 标注规则

本小节分别介绍事件事实性相关信息的标注规则。在例子中出现的下标意义如下：“s“、“p”、“e”、“n”、“d”和“c”分别表示事件源、事件选择谓词、触发词、否定词、程度词和从句，后面跟着的数字表示序号。有属性的事实性相关信息，其属性表示在该词后面"[]"中。

### 2.3.1 标注事件选择性谓词和事件源

在 ACE 2005 中文语料库中，文档的媒体来源包含在文档名字中，如新华社的新闻稿文档名以“XIN”开头，这类媒体源识别简单。作者源通常以固定形式出现在新闻题材或者博客文档中，如“我是徐亚文，欢迎您继续收听新闻”中的“徐亚文”。本节主要介绍直接源的标注，它通常和事件选择谓词一起出现，因此将这两者的标注一起介绍。

一般情况下，事件选择谓词是动词。除此之外，也会出现名词情况。

1）事件选择谓词为动词时，选择谓词通常为事件触发词的上层谓词，事件在谓词的论元中被提及。直接源是谓词的主语。如例 E8-E9 所示。

*E8：当地政府$_{s0}$指责其暴行并加派$_{p0}$[级别：确定]警力前往$_{e0}$边境加强保护。*

注意：在中文中主语缺省是一个普遍的现象。在这种情况下，事件选择谓词和事件触发词连着出现，位于触发词前。

*E9：明天，他$_{s0}$计划$_{p0}$[级别：可能]（他）到$_{e0}$上海去。*

2）事件选择谓词为名词时，通常是以一个复句中的分句形式出现。如例 E10 所示。

*E10：根据被害人$_{s0}$的证词$_{p0}$[级别：确定]，大约是半夜十二点回到$_{e0}$家的。*

如果将事件选择谓词按语义分类，可以分成以下几大类：

发表型谓词：例如：宣称，告诉，透露等等。当然也包括非口头式的，如撰写、公布，发表等等。

认知型谓词：所谓认知型谓词是指对应的事件源原本对这个事件有了解（例如：知道，了解，懂得，记得），或者发现了某件事（例如：发现），或者忘记某件事（例如：忘记，遗忘），或者对某件事表示认可（例如：赞成，支持，承认，接受等等）。

意见型谓词：比较典型的意见型谓词有：建议，认为，觉得，猜测，考虑等等。就是事件源对之所持有的态度意见想法或者建议。

疑问型谓词：事件源对该事件怀有疑问。比较典型的有：怀疑，询问，好奇等等。

感觉型谓词：如看到，听到，感觉到等等。

推理型谓词：这一类谓词可以表示一种推理或者推断的过程，如：推断，得出（结论），引出等等。

生理型谓词：生理型谓词是指事件源对该事件发生的反应，一般有：愧疚，悔恨，欣喜，感激等等。

证明型谓词：如证明，显示，解释等等。

对应的直接事件源也就是某言论的发表者，某意见或者疑问的持有者，有某种感觉或者反映的载体等等。

2.3.2 标注否定词

否定词表示否定意义的词，一般是修饰动词的副词，有时也有动词情形。在标注否定词的过程，要关注否定词的作用范围，在本文中只标注本事件在其作用范围内的否定词，否则一律不标注。在例 E11 和 E12 中，否定词的范围用“[]”表示。

1）否定词为副词

在动词前，修饰动词，作用范围为该动词在内的子句。

*E11:因为天气原因，[会议$_{e0}$并没有$_{n0}$如期举行]。*

2）否定词为动词。否定词为动词时，其作用范围通常为该否定词所在的子句。

*E12：匕首型刀刃如果不用应该及时收好，[避免$_{n0}$被利器割伤$_{e0}$]。*

2.3.3 标注程度词

程度词表示事件的确定性程度，主要有副词和助词构成，极少情况也会出现动词和形容词等。程度词有时态和级别属性，但不是所有的程度词都同时具备这两个属性，有的只有其中之一，因此，把程度词分为三类分别讨论。

1）时间程度词

时间程度词也就是纯时态词，一般为表示时间的副词，例如：刚刚、将、日前等等。下例 E13 中“将”是一个表示将来的时间程度词，它并不表示事件确定性的程度，其级别属性为无。

*E13：越南国家主席陈德良将$_{d0}$[时态：将来，级别：无]在 12 月 15 号到 29 号访问$_{e0}$中国大陆。*

2）级别程度词

级别程度词是表示语气的副词，位于动词前修饰动词，表示事件事实性的确定性的程度。例如：可能，大约，大概，必须，的确等等。例 E14 中“可能”是一个表示可能的程度词。

*E14：谭梅清可能$_{d0}$[时态：无，级别：可能]离开$_{e0}$上海了。*

3）混合程度词

混合程度词比较特殊，既表示程度又表示时间。按词性分类可以是助词、动词和副词。

混合程度词可以是助词，接在动词后，表示结果，如：（涌进）了，（救）下等等。例 E15 中“了”是一个表示过去、确定的混合程度词。

*E15：民众由各地涌进 $_{e0}$ 了 $_{d0}$[时态：过去，级别：确定]市区。*

混合程度词可以是动词，一般以能愿动词为主，如：要，能等等。例 E16 中“要”是能愿动词，表示事件源“她”的意愿，因此其时态属性为将来，级别属性为可能。

*E16：她宣称是要 $_{d0}$[时态：将来，级别：可能]带儿子到 $_{e0}$ 宾州的一所军校就读。*

混合程度词也可以是副词，在动词前修饰动词。例 E17 中“已经”，作为副词修饰动词“移居”，其时态属性为过去，级别属性为确定。

*E17：彭春燕已经 $_{d0}$[时态：过去，级别：确定]移居 $_{e0}$ 美国了。*

2.3.4 标注从句

从句的识别相对比较简单，本文只标注显示从句，只要根据显示连接词标注即可。如下例 E18 中“如果”就是条件从句的显式连接词。

*E18：如果你赶在国庆之前回来 $_{c0}$，我就不走 $_{e0}$。（显式条件从句）*

## 3 语料标注工具和标记

为了提高标注效率，我们开发了一个标注工具。在标注过程中，标注人员只需在“全文”栏中选中需要标注的内容，然后点击相应的标签按钮即可。例如：标注事件源，首先在“全文”栏中选中事件源（如“新华社”），然后点击“源”标签按钮，在“标注结果”框中就会显示相应标注的结果。

在本文中，标注工作在 ACE 2005 中文语料库的基础上展开，标记方式遵循 ACE 2005 语料库的标记方法。标注结果以 XML 标签的形式存储在 ACE 2005 原有的 apf 文档中。

其中，<facutality>表示事实性信息的根结点，<event_selecting_predicate>、<txt_predicate>、<source>、<txt_source>、<text_source>、<clause>、<negative_word>、<degree>作为<facutality>的子标签分别表示事件选择谓词、文章级谓词、直接源、作者源、媒体源、从句、否定词、程度词的具体标注内容，例 19 是一个标注示例。

*E19：*

```
<factuality>//中文事件的事实性
<degree LEVEL="无" TENSE="过去">//程度词
<charseq START="170" END="173">下午5点</charseq>
    </degree>
<event_selecting_predicate LEVEL="确定">//事件选择谓词
<charseq START="167" END="168">宣布</charseq>
    </event_selecting_predicate>
<source ID="CBS20001101.1000.0000-E11-23">//直接源
<charseq START="159" END="165">香港机场管理局</charseq>
    </source>
<text_source>//媒体源
<charseq>CBS</charseq>
    </text_source>
</factuality>
```

## 4 数据统计与分析

事件事实性的标注工作只是我们工作的第一步，未来我们希望能够实现自动抽取事件的事实性信息，并利用事实性信息来提高事件抽取系统的性能。因此，本小节中，我们对此次标注的语料库进行一系列的统计和分析，希望为我们即将开展的工作提供基础性的指导，便于我们分析具体的语言现象。

目前本文已经标注的语料有 297 篇文档，标注对象是 ACE 2005 中文语料库中 Movement 事件的事实性。下表 2 是五类事实性相关信息的标注数量，其中事件选择谓词不包括文档级别的谓词，事件源的个数不包括作者源和媒体源的个数。从表 2 可以发现，事件选择谓词和事件源的数量相当，因为它们通常结对出现，极少数情况会省略源。另外，文档级别的谓词一般出现在新闻题材的文章中，其标注情况如图 1a)所示，可以发现其出现的项相对集中稳定。作者源的标注，根据其题材的不同而不同。其中新闻稿和广播的作者源以标注具体的新闻社或者媒体平台为主，没有则标注记者名。博客部分标注发布者的名字，一般为网络名。其统计情况如下图 1b)所示。

表 4-2 中文事件事实性信息语料库具体标注项统计

| 标注项 | 事件选择谓词 | 事件源 | 否定词 | 程度词 | 从句 |
|---|---|---|---|---|---|
| 标注个数 | 174 | 169 | 39 | 235 | 11 |
| 含该项句子数占比 | 23.0% | 22.3% | 5.1% | 31.0% | 1.5% |

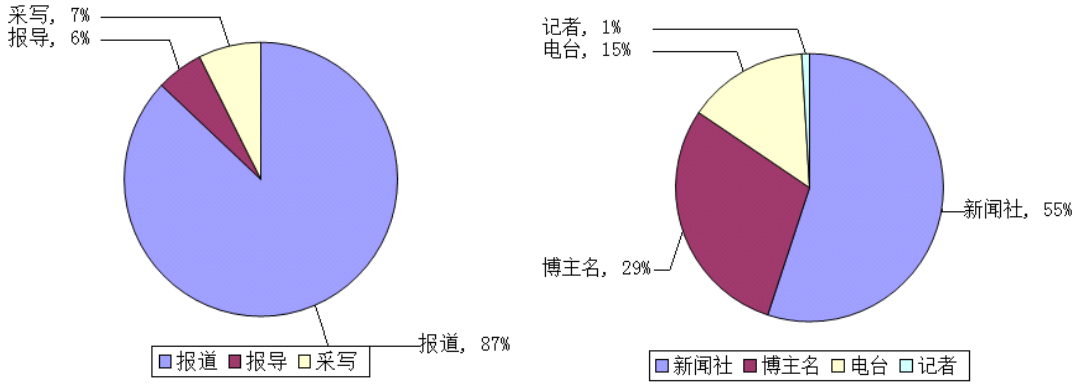


图 1 a)文档级别的谓词标注情况统计图; b)作者源的标注情况统计图

图 2 是程度词的时态属性和级别属性标注结果的统计图。图 2a)中，过去时态的程度词最多，占总数一半以上；图 2b)中，级别为确定的程度词最多，占总数一半以上；级别程度词和时间程度词，分别占总数的 13%和 28%。表 3 是程度词的时态和级别属性对标注结果的统计情况。从表 3 发现，(过去，可能)、(过去，不确定)、(现在，可能)和(现在，不确定)这四种属性对出现次数均为 0，这符合常识推理，因为一般过去和现在发生的事件都是确定的，这既表现在句子的词语之间的相互呼应和支撑，如“已经...了”等。这也表现在具体的词语的语义上，但是如果要表示过去的某件事并不能确定的语义情况，只能通过句子的不同表达方法来表达，如“可能已经...了”等。通过一个词语既要表现过去或者现在的时态，又要表示可能或不确定的语义，现实中是不存在的。表 3 的最后一行和最后一列是时间程度词和级别程度词的个数，其中表示将来的时间程度词最多，表示可能的级别程度词最多。

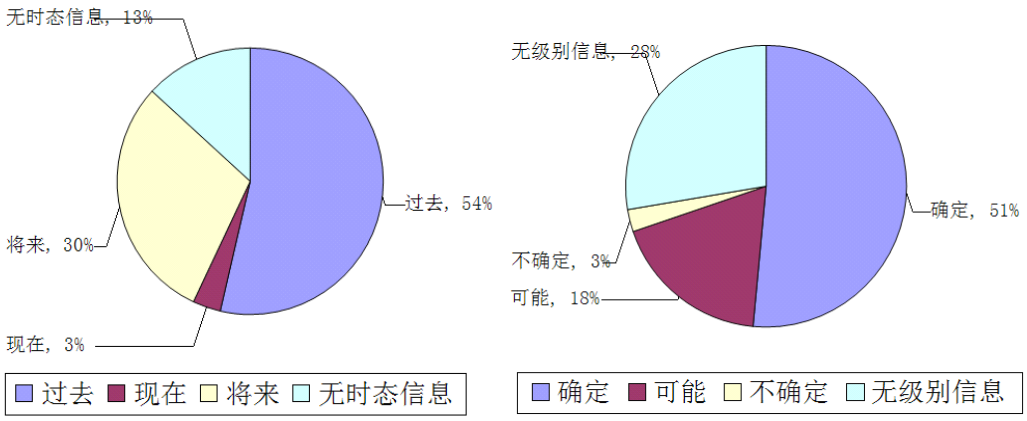


图 2 a)程度词的时态属性标注情况统计图; b)程度词的级别属性标注情况统计图

表 4-3 程度词的属性对标注情况统计

| (时态，级别) | 个数 | (时态，级别) | 个数 | (时态，级别) | 个数 | (时态，级别) | 个数 |
|---|---|---|---|---|---|---|---|
| (过去，确定) | 110 | (现在，确定) | 1 | (将来，确定) | 5 | (无，确定) | 5 |
| (过去，可能) | 0 | (现在，可能) | 0 | (将来，可能) | 24 | (无，可能) | 19 |
| (过去，不确定) | 0 | (现在，不确定) | 0 | (将来，不确定) | 0 | (无，不确定) | 6 |
| (过去，无) | 16 | (现在，无) | 7 | (将来，无) | 41 | (无，无) | 1 |

表 4 是谓词的级别属性统计情况表，其中级别为确定的事件选择谓词最多，其次是可能，表示不确定的谓词个数为 0。表 5 是事实性结果的统计表，确定发生的事件最多，达到了 66%，其次是可能发生事件，达到了 20%。

表 4-4 谓词的级别属性标注情况统计

| 属性-值 | 级别-确定 | 级别-可能 | 级别-不确定 |
|---|---|---|---|
| 个数 | 145 | 29 | 0 |
| 百分比 | 86% | 14% | 0% |

表 4-5 事件的事实性标注结果统计

| 事实性 | 当然发生（C+） | 可能发生（P+） | 可能不发生（P-） | 当然不发生（C-） | 不确定（U） |
|---|---|---|---|---|---|
| 事件个数 | 498 | 150 | 26 | 18 | 9 |
| 占比 | 66% | 20% | 3% | 2% | 1% |

最后，我们采用 kappa 值作为衡量语料标注一致性的指标。在这里，我们利用两位标注者的标注结果进行一致性计算，由于结果 kappa 值很理想，因此，没有另外再请第三位标注者进行标注。其中，五项事实性相关信息和三维体系的一致性分开独立计算，互不影响。在计算一致性过程中，对于从句采用模糊匹配。例如：“为了和小三结婚”这个目的从句，标注“为了”或者“为了和小三结婚”都可。其余的采用完全匹配的方式，只有当两位标注者标注的内容完全一致时，才认为两位标注者一致同意该实例的标注。具体一致性统计结果，如下表 6、7 所示。最后，还统计并且计算了两位标注者对事件事实性结果的一致度，大约为 68.0%。

从表 6 中可以看出，否定词的 kappa 值不高，主要原因是进行标注工作的两位标注者，一位是标准规则的制定者，另一位标注者则是其他项目组的成员，对本次标注领域并没有特别深刻的研究，因此在否定词标注过程中，只标注了“显式否定词”，如“不”，“没有”等等，对于一些具有否定性推理意义的词并没有标注，如“原定”，“押后”等等。

表 4-6 事件事实性相关项的 kappa 值统计

| 项 | 事件选择谓词 | 事件源 | 否定词 | 程度词 | 从句 |
|---|---|---|---|---|---|
| Kappa 值 | 62.1% | 64.2% | 61.0% | 70.6% | 72.3% |

表 4-7 三维体系 kappa 值统计

| 项 | Degree | Polarity | Genericity | Tense |
|---|---|---|---|---|
| Kappa 值 | 59.3% | 76.0% | 59.9% | 69.2% |

## 5 总结

本文首先介绍了中文事件事实性信息语料库构建的必要性和当前存在的问题，然后介绍了中文事件事实性和一组影响事实性的具体项——事实性相关信息。本文重点介绍了这一组事实性相关信息的标注规则，并且在 ACE 2005 中文语料库的基础上，标注了 Movement 事件的事实性。最后对目前完成标注的小型语料库进行统计分析，为下一步工作的开展提供基础。从统计结果发现，确定发生的事件比率最高，超过总数的一半。Movement 事件事实性的标注是标注工作的一部分，标注其他类型事件事实性以扩大语料库规模是下一阶段的工作，将来我们希望可以自动抽取事件的事实性相关信息，并利用事实性信息来提高事件抽取系统的性能。

## 参考文献


[1] Saur R., Pustejovsky J. FactBank: A Corpus Annotated with Event Factuality. Language Resources and Evaluation,2009,43(3):227-268.

[2] Saurí R., Pustejovsky J. Are you sure that this happened? Assessing the Factuality Degree of Events in Text. Computational Linguistics,2012,38(2):261-299.

[3] Saurí R. FactBank1.0 Annotation Guidelines,2008.

[4] Vincze V., Szarvas G., Farkas R., Mora G., Csirik J. The BioScope Corpus: BioMedical Texts Annotated for Uncertainty, Negation and Their Scopes. BMC Bioinformatics, 2008.

[5] Riza T., Sophia A. Building a Coreference-Annotated Corpus from the Domain of Biochemistry. In Proceedings of the 2011 Workshop on Biomedical Natural Language Processing,ACL-HLT, 2011:83-91.

[6] Vincze V. Speculation and Negation Annotation in Natural Language Texts: What the Case of BioScope might (not) Reveal. In the Proceedings of the Workshop on Negation and Speculation in Natural Language Processing. Association for Computational Linguistics,2010:28-31.

[7] Kim J., Tomoko O., Junichi T. Corpus Annotation for Mining Biomedical Events from Literature. BMC Bioinformatics,2008.

[8] Ohta T., Tateisi Y., Kim J. The GENIA Corpus: An Annotated Research Abstract Corpus in Molecular Biology Domain. In the Proceedings of the second international conference on Human Language Technology Research,2002:82-86.

[9] Rubin, V. L. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In Proceedings of the NAACL-HLT,2007:141-144.

[10] Saurí R., Verhagen M., Pustejovsky, J. Annotating and Recognizing Event Modality in Text. In Proceedings of the 19th International FLAIRS Conference, FLAIRS,2006.