构建大规模的汉语事件知识库

周 强[1], 王俊俊[2], 陈丽欧[2]

[1]清华大学 信息技术研究院语音和语言技术中心

[1]清华信息科学与技术国家实验室

[2]清华大学 计算机科学与技术系，北京 100084

E-mail: zq-lxd@tsinghua.edu.cn; jears06@gmail.com; chouou@foxmail.com

摘 要：随着互联网的迅猛发展，大量的信息以文本的形式快速涌现。对海量文本进行信息的深度挖掘离不开高质量的事件内容分析技术，而这些技术的开发又需要高质量的事件语义标注资源支持。本文提出了一个构建大规模汉语事件知识库的可行方案。实验证明，我们的方案能很好地解决事件知识库“可操作性，可计算性，可扩展性”问题。既可以在较少的投入条件下，“小而精”地解剖一个局部问题，又可以方便扩展到更大的领域和更多的应用中。

关键词：事件内容分析；事件语义标注资源；汉语事件知识库

Building a Large-scale Chinese Event Knowledge Base

Zhou Qiang[1], Wang Junjun[2], Chen Liou[2]

[1] Center for Speech and Language Technology, Research Institute of Information Technology

[1] Tsinghua National Laboratory of Information Science and Technology

[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084

E-mail: zq-lxd@tsinghua.edu.cn; jears06@gmail.com; chouou@foxmail.com

Abstract: With the rapid development of Internet, large quantities of information appear in the form of e-text. Mining deep information in mass texts is inseparable from high-quality event content analysis technology. And these technologies need the support of high-quality annotated resources of event semantics. This paper presents a solution on how to build a large-scale Chinese Event Knowledge Base. Some primitive experiments have shown that our approach can deal with the problems of operability, computability and scalability in developing a large-scale semantic knowledge base. Under limited conditions, we can describe the event knowledge in a small local domain easily and excellently. Also, it can be easily extended to larger areas and more applications.

Keywords: event analysis; event annotation; event knowledge base

1 引言

随着互联网的迅猛发展，大量的信息以文本的形式快速涌现。如何从海量的文本中准确抽取到所需要的信息，已经成为研究的热点问题。

对海量文本进行信息的深度挖掘离不开高质量的事件内容分析技术，而这些技术的开发又需要高质量的事件语义标注资源支持。近几年来，英语方面陆续启动了多个大规模的事件语义资源开发项目，如 FrameNet[1]、OntoNotes[2]等，它们分别从不同角度对英语真实文本句子中的事件语义信息进行了深度标注。在这些项目的推动下，事件语义资源的开发取得了长足的进展和较为丰硕的成果。相对而言，汉语的事件语义资源开发还很薄弱，需要进行大量工作。

针对汉语的研究现状，结合汉语自身的特点，我们设计并实现了一个针对汉语客观事件的句法、语义和概念描述知识库——汉语事件知识库。该项目得到了国家 863 计划课题的支持，由北京大学、鲁东大学和清华大学协作开发完成。

在一个统一的设计框架下，我们将相关事件知识描述拆分成五个子库，包括两个静态库、两个动态库以及一个用于在两大知识库之间建立联系的动词义项对齐知识库。五个子库相互配合，互为补充，为汉语文本的事件内容分析提供语义资源的支持。初步的实验结果显示，我们的方案能

很好地解决事件知识库的“可操作性，可计算性，可扩展性”问题。在较少的投入条件下，“小而精”地解剖一个局部问题，并可以很容易地扩展到更大的领域和更多的应用中。

在此基础上，我们进一步分析了各子库的内在关系，提出构建集成事件知识库的设想，以挖掘知识库中的隐含信息，建立统一的事件描述体系，为开发更好的汉语事件计算平台提供条件。

本文其余部分安排如下：第二章介绍汉语事件知识库的整体框架、各子库的详细内容及开发现状；第三章结合实例对汉语事件知识库的结构进行进一步展示，并着重分析各子库间的内在联系；第四章提出构建集成的大规模事件知识库的设想；第五章分析了在事件语义资源方面现有的相关研究成果；第六章是对现有工作的总结和对未来工作的展望。

2 汉语事件知识库开发

2.1 总体框架

在汉语事件知识库开发过程中，我们提出了静态知识库和动态标注库相结合的构建路线，从两个不同角度对特定事件内容进行深入描述和知识挖掘：静态库汇集了大量的语言学专家描写知识，动态库提供了丰富的客观事件标注实例。

在静态知识库方面，我们设计了情境网络和词汇知识库两个子库。前者侧重从语义概念层面对不同事件、关系和状态进行细致描述，形成概念层面进行知识推理和语义计算的基础知识单元；后者侧重从词汇语义层面对不同词语内部隐含的句法语义分布信息进行描述，以便建立真实文本描述实例与词汇语义知识库之间的内在联系。通过以上两个静态库，我们可以建立从表层的词汇描述形式到深层的情境概念表达之间的联系通道，为实现对表层文本反映的深层客观事件内容的准确分析和相关知识推理提供支持。

在动态标注库方面，我们设计了目标动词义项标注库和事件描述块句法语义标注库两个子库，分别从目标动词义项和事件描述块句法语义两个层面对真实文本中的事件内容进行挖掘。通过对真实文本句子中事件目标动词义项和事件描述块的句法语义信息的准确标注，形成了大规模的客观事件内容描述实例，为相应语义计算工具的知识获取和统计建模提供有力支持。

为了有效地建立起静态库和动态库之间的联系，我们设计了事件目标动词义项对齐知识库。通过人工标注，实现各个语义词典之间的义项对应，明确各个语义词典提供的事件框架之间的角色对应关系。以这个对齐知识库为中间桥梁，可以方便地建立起两大知识库之间的信息联动。

图 1 显示了事件知识库开发的总体结构。

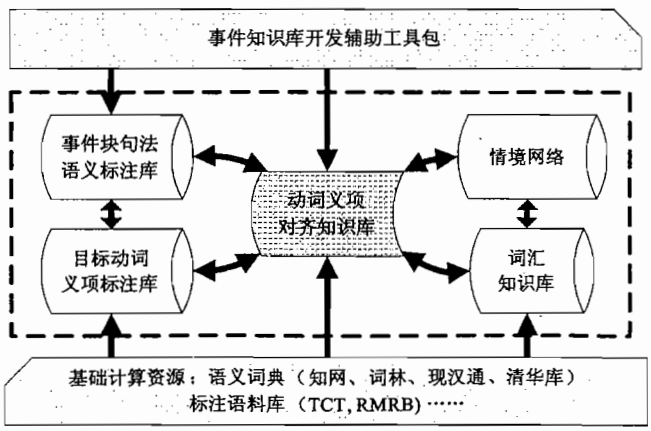

图 1　事件知识库开发总体结构

在一个统一的设计框架下，相关事件知识描述被拆分成 5 个既相互独立又存在内在信息联系的知识子库。经过有效拆分和信息联动，一方面细化了工作的粒度，便于分工合作，另一方面又增强了信息的可靠性和丰富性，提高了描述的质量。

2.2 分库基本内容介绍

从我们关注的特定事件类型出发，各个子库分别从不同的角度对事件相关知识进行描述。

2.2.1 情境网络

情境网络描述体系[3]从概念语义层面对事件进行描述，通过不同的情境关系，建立起这些情境反映的事件内容之间的内在联系，形成概念层面进行知识推理和语义计算的基础知识单元。其构建过程主要包括情境的划分，情境网络的构建，以及定义词汇的确立。其描述核心是通过相关信息抽象形成的情境表达式，并通过“相关情境”、“定义词汇”等建立库内和库间的信息联系。

为了便于进行知识推理和语义计算，在情境的划分过程中，我们力图保证情境概念描述的概括性和全面性。同时，为了便于人工分析把握，在情境网络的构建过程中，我们控制每个主题网络的规模，并限制定义词汇的数量。在确定各情境的定义词汇时，我们尽可能地遵循以下原则：

a) 一个特定情境的所有定义词汇具有相同的核心参量，参量之间的句法语义关系相同。

b) 子情境与子情境之间，定义词汇成对立互补分布。对于可能激活不同情境的动词，将其拆分为不同的义项，归入对应的情境中。

以“领属变化”类事件为例，我们将相关事件拆分成“失去”、“获得”、“转让”、“商品交易”、“赊购”、“借还”、“租赁”等主题情境网络，每个主题情境网络包含 7~10 情境，每个具体情境中包含若干定义词汇，同时主题情境网络之间也存在一定的联系。

2.2.2 词汇知识库

词汇知识库从词汇语义层面对不同词语的句法语义分布信息进行描述。对于可能激活不同情境的动词，在词汇知识库中都被拆分成不同的动词义项，分别进行句法语义分布的描述。从而保持了两个静态库的一致性，更好地反映各情境事件的区别和联系。

词汇知识库的描述核心是相关事件义项的语义论旨角色和句法配置模式，这是静态知识库与真实文本标注实例之间建立联系的重要桥梁。同时，词汇知识库还通过“情境定义”、“义项描述”、“参量锚定”等信息与情境网络、动词义项对齐知识库建立了联系通道。

2.2.3 目标动词义项标注库

目标动词义项标注库精选人民日报标注库、清华树库[4]真实文本句子，对句子中的单义或多义目标动词进行词义的区分和对应，从而实现对目标动词全方位、多实例的刻画和描述。

在目标动词义项标注方面，我们选择了现有的三个典型语义词典：知网[5]、词林[6]和现汉[7,8]。它们分别采用了义原表达式、同义词集合和自然释义三种方式来描述事件意义。根据真实文本句子中各个目标动词出现的不同语境，分别选择上面三个词典中的合适义项描述，可以形成多个词典对齐的义项标注信息[9]。这样，一方面可以充分利用三个词典中的义项描述信息形成信息互补的完整事件内容描述；另一方面，也可以利用相关标注提供的不同语义词典计算入口，集成各个词典的计算能力实现我们需要的事件语义计算任务。

2.2.4 事件块句法语义标注库

事件块语义句法标注库的标注文本选择与目标动词义项标注库相同。

事件描述块的句法语义信息标注，主要是在目标动词控制的事件描述小句中，进一步确定该目标动词所反映事件情境的各个描述块，并对其进行句法语义信息标注，包括：确定块边界、标

注句法功能和成分标记、确定各个块的中心词位置、标注合适的语义角色标记等。另外，还对代词指代和角色省略问题进行了特殊处理，通过寻找和标注事件描述小句块外部的对应信息，保证了相关事件内容描述的完整性[10]。

2.2.5 动词义项对齐知识库

事件目标动词义项对齐知识库是各个子库之间联络的核心和枢纽。我们从静态知识库和动态标注库中的各动词出发，依托知网、词林、现汉三大语义词典，通过人工标注，明确三个语义词典中动词各义项之间存在的对应关系。进一步，对于语义词典提供的事件框架，联系情境网络中的“参量锚定”和词汇知识库中的“论旨角色”，以动态标注为参考，确定其角色对应关系，搭建起静态库和动态库之间信息通道。

2.3 目前开发状况

开发大规模的事件语义资源需要巨大的工作量，消耗大量的人力财力。在有限的资源限制下，我们的方案可以针对关注的特定事件类型，建立一致、系统的知识架构，提供准确、全面、且相互融会贯通的语义资源，“小而精”地解决一个特定问题。这已在“存在拥有类”事件知识库开发工程中得到了可行性和有效性验证，可以方便地推广到其他类似的事件知识库开发过程中，从而很好地解决事件知识库的“可操作性，可计算性，可扩展性”问题。

事件知识库的开发现状参见表 1。

表 1 事件知识库开发现状

事件知识库	开发现状
静态知识库	广义拥有关系：49 个情境，736 个词语义项 时空变化状态：23 个情境，812 个词语义项
动态标注库	119 个多义动词、479 个单义动词 10 万个真实文本句子中的事件内容标注
动词义项对齐库	1669 个描述记录的义项和事件框架对齐信息

3 实例分析及子库内在联系挖掘

本章以目标动词“租赁”作为切入点，通过详尽的实例分析，对事件知识库的结构进行进一步的展示，并着重分析各子库间的内在联系。

与“租赁”相关的主题情境网络（租赁情境子网络）如图 2：

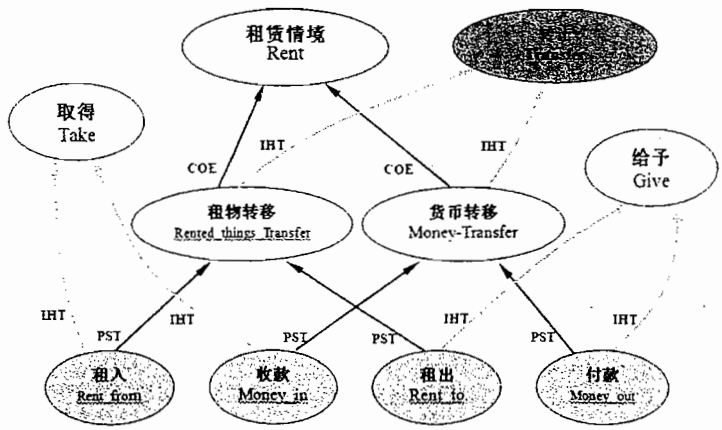

图 2 租赁情境子网络

一个租赁事件一般包括同时发生的两个子事件：租物转移和货币转移。它们的下一层又对应若干具体的动作事件，如租入、租出、收款、付款，而这些情境又分别属于取得、给予、转让等情境的范畴。这样，我们将每个事件和与其相关的其他事件通过情境网络联系起来，通过情境网络中对相应情境关系的界定和描述[3]，为相应的事件分析和知识推理提供了依据。

动词“租赁”有两个含义：租出和租入。按照之前的约定，我们将其拆分为两个义项，“租赁1”和“租赁2”，分别对应了租出和租入情境。以“租入”情境和“租赁2”义项为例，各个子库的信息描述单元及相互之间的对应关系如图3。

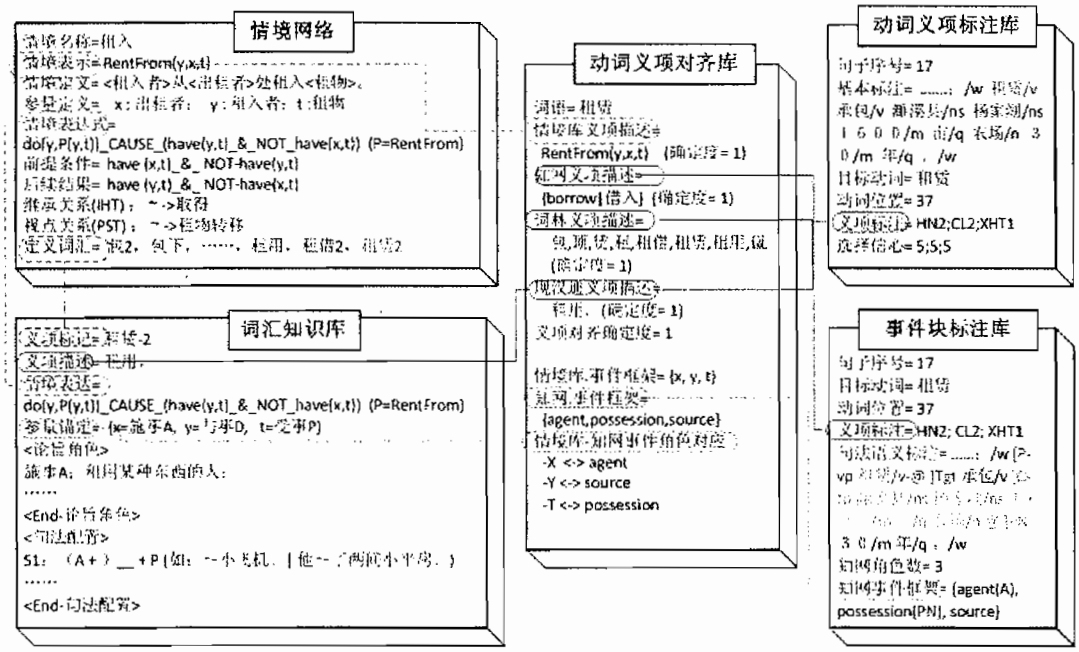

图3　事件知识库各子库信息描述单元及相互对应关系

情境概念单元的描述主要包括三个部分：情境的完整描述，包括情境名称、情境表示、情境定义、参量定义、情境表达式、该情境发生的前提条件和后续结果等；情境与相关情境的关系描述，它们形成了情境网络的推理关系；情境的定义词汇信息。其中，“情境表示”可以直接联系到动词义项库中的“情境库义项描述”，而“情境表达式”和“定义词汇”项可以与相应的词汇知识库记录进行双向连接。

词汇知识库的义项描述单元包括如下内容：义项描述、情境表达、参量锚定 、论旨角色和句法配置。其中，“义项描述”可以与动词义项库中的现汉通义项建立对应关系；“参量锚定”可以建立情境参量和论旨角色之间的内在联系，同时又与动词义项对齐库中的事件角色对应关系互为呼应，从而建立起相应信息与知网事件角色之间的内在联系；“句法配置”描述了文本句子中不同语义论旨角色的典型配位形式，提供句法语义连接信息。

目标动词义项标注库汇集了大量以“租赁”为目标动词的真实文本句子实例。对每个目标动词，给出了在该实例文本句子语境下的知网、词林和现汉通的义项选择和人工标注信心信息。其中“义项标注”可以与动词义项对齐库中的知网、词林、现汉通义项描述建立联系通道。

事件块语义句法标注库对句子中目标动词控制的各个事件块，分别给出了句法功能(S,P,O)、句法成分(np, vp, tp)和语义角色标注(A,PN)，并用“@”符号标注出了每个事件块的语义中心词信息。“义项标注”和“知网事件框架”与动词义项对齐库中建立了对应关系。

动词义项对齐知识库主要包括该义项在不同词典中的情境意义描述和各个主要事件框架的角色对应信息描述，以实现方便的建立起动态库和静态库之间的内在联系。其中，“情境库义项描述”、“词林义项描述”、“知网.事件框架”分别与之前所述各记录信息呼应，在各个子库之间搭建起了信息通道，从而得到了图 1 所示的事件知识库互连互动框架体系。

4 集成事件知识库开发设想

事件知识库是一个相互关联的有机整体，但是这种关联性隐含在各个子库中，不够集中和直观。在开发过程中，子库的拆分降低了分析和标注的难度，但在实际运用中，我们更关注其易用性和语义计算性能。而且，在人工合作分析标注的过程中，难免出现子库间的不一致、不同步。为了更有效地发挥事件知识库的研究和应用价值，需要在信息的集成和统一的事件内容计算平台开发方面进行更深入的工作。由此，我们进一步提出了集成事件知识库的开发设想。

首先按照各子库给出的事件描述深度的不同，将它们重新组织成三个基本事件知识库：1) 情境描述库；2) 事件描述库；3) 标注句子库。其中，情境描述库侧重对某类事件的内容抽象和关系挖掘，形成可以进行初步知识推理的情境网络，其基本信息来自现有的情境网络描述库；事件描述库侧重对某个事件的内容描述，通过建立各个语义资源的义项描述和事件框架之间的内在联系，提供各个语义资源之间的计算入口，其基本信息通过融合现有的词汇知识库和动词义项对齐库信息得到。标注句子库侧重对真实文本句子中某个事件内容的信息标注，包括事件目标动词的义项标注和该目标动词控制的事件块的句法语义标注等，其基本信息来自现有的两个动态标注库。

其次，通过对低层次资源的数据汇总分析，可以为高层次资源提供更多更详细的人工标注互补分析数据，为进一步改进相关资源的计算能力提供支持。

目前我们已经完成情境描述库的构建，并且检查和明确了情境描述库和词汇知识库之间的双向联系。其他的集成和对齐工作正在进行中。

5 相关研究工作综述

近年来，国外多个构建大规模事件语义资源的项目陆续启动，本章将对其中有代表性的项目进行介绍。

ACE(Automatic Content Extraction)[11]项目的目标是研究文档内容的抽取技术，包括实体、关系、事件等，主要关注网络上的专线新闻、网络日志等 6 个领域，提供英文、中文、阿拉伯文三个语种的训练语料，2007 年增加了西班牙语。ACE 语料以篇章为单位，详细标注了底层的标准实体、时间、值的信息。ACE05 提供了英、中、阿三种语言 300K 的训练库和 50K 的测试库。

OntoNotes[2]的目的在于构建大规模的跨领域标注语料库，涵盖英文、中文、阿拉伯文三种语言的新闻、电话对话、网络日志、脱口秀等文本。OntoNotes 语料库中标注了语言的结构信息（句法树和谓词论元结构）和浅层语义信息（动词、名词的词义及共指关系）。最新发布的 4.0 版本包含 300K 的阿拉伯语料，800K 的汉语语料，以及 1300K 的英文语料。

Berkeley FrameNet[1]以框架语义为标注的理论基础，试图发现核心动词(LU)和它周围各框架元素(FE)之间的搭配关系，从而归纳出知识的语义表示方法，进而集结各框架构成 FrameNet 网络。FrameNet 的语料来源于英国国家语料库，每个句子都标注了目标谓词和其语义角色、该角色句法层面的短语类型以及句法功能。最新数据显示，FrameNet 已包含 11,600 个词条，960 个事件框架和 150,000 个标注句子。

Propbank[12]是集语义词典和标注语料库于一身的论元角色语义知识库。它以动词词典为标注基础，以 Penn TreebankII 为标注底层，以动词的论元角色为标注对象。PropBank 为超过 3300 个动词建立了 4500 个框架，并在中文 Treebank 基础上，构建了 500K 的中文 PropBank 语料[13]。

TimeML 项目的语料资源主要是 TimeBank[14]。TimeBank 主要来自 Wall Street Journal 和 New York Times 的新闻文章，根据 TimeML 的标准，详细标注了事件、时间表达式以及它们之间的时序关系。到目前为止，TimeBank 的最新版本为 1.2，共包含 183 篇新闻文章，7935 个事件。

可以看出，大多数的事件语义资源开发将侧重点放在真实文本句子的标注上，FrameNet 从框架语义学出发，试图归纳知识的语义表示方法，这与我们的做法很类似，但还是有所不同。我们的汉语事件知识库从静态知识库与动态标注库两个角度对事件内容信息进行挖掘和描述，且所有五个子库是在一个统一的设计框架下展开，因此可以关注特定事件类型，有针对性地以较少的代价“小而精”地解决问题，同时又具有非常好的可扩展性。

6 结语

近年来，在多个项目的推动下，事件语义资源的开发取得了长足的进展和较为丰硕的成果。相比之下，国内对于汉语事件语义资源的开发明显薄弱不足，所以，探索大规模的汉语事件知识库的开发和建设有其紧迫性和必要性，以及重大的应用价值和长远意义。

我们针对汉语的研究现状，结合汉语自身的特点，设计和开发了“汉语事件知识库”。在一个统一的设计框架下，相关事件知识描述被拆分成 5 个既相互独立又存在内在信息联系的知识子库。通过各个子库之间的相互配合和信息联动，可以提高各自的描述质量。在此基础上，我们又进一步提出开发集成的事件知识库的设想，希望对推动汉语文本自动分析技术的发展有所帮助。

7 致谢

本课题得到了国家 863 项目（编号：2007AA01Z173）、国家自然科学基金（编号：60873173）、Tsinghua-Intel 合作研究项目的支持。情境网络和词汇知识库由北京大学袁毓林教授领导的研究小组完成，目标动词义项和事件块句法语义标注库由鲁东大学亢世勇教授领导的研究小组完成。在此一并致谢。

参 考 文 献

[1] Ruppenhofer J, Ellsworth M, Petruck M R L, et al. FrameNet II: Extended Theory and Practice. http://framenet.icsi.berkeley.edu/.

[2] Weischedel R, Pradhan S, Ramshaw L, et al. OntoNotes Release 4.0. http://www.bbn.com/NLP/OntoNotes/.

[3] 北京大学汉语语言学研究中心. “广义拥有”与“领属变化”情境网络描述体系[R]. 技术报告. 2009.

[4] 周强. 汉语句法树库标注体系[J]. 中文信息学报. 2004, 18(4): 1-8.

[5] 董振东, 董强. 知网. http://www.keenage.com/.

[6] 梅家驹, 竺一鸣, 高蕴琦等编. 同义词词林[G]. 上海辞书出版社, 1983.

[7] 中国社科院语言研究所词典编辑室. 现代汉语词典（修订本）[G]. 商务印书馆, 1996.

[8] 中国人民大学语言文字研究所. 现代汉语通用字典[G]. 外语教学与研究出版社, 1987.

[9] 鲁东大学中文信息处理研究所. 目标动词义项标注规范6.0[R]. 技术报告, 2009.

[10] 鲁东大学中文信息处理研究所. 事件描述块句法语义标注规范6.0[R]. 技术报告, 2009.

[11] Doddington G, Mitchell A, Przybocki M, et al. The automatic content extraction (ace) program-tasks, data, and evaluation[C]. In: Proceedings of LREC. 2004. 837-840.

[12] Palmer M, Gildea D, Kingsbury P. The proposition bank: A corpus annotated with semantic roles[J]. Computational Linguistics. 2005, 31(1): 71-106.

[13] Xue N, Xia F, Chiou F D, et al. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus[J]. Natural Language Engineering. 2005, 11(02): 207-238.

[14] Pustejovsky J, Hanks P, Saurı R, et al. The timebank corpus[C]. In: Proceedings of Corpus Linguistics 2003. 2003. 647-656.