事件词驱动的文本事件信息结构初探*

曾青青，杨尔弘
北京语言大学 应用语言学研究所，北京 100083
E-mail: qing8612@sina.com; yerhong@126.com

摘　要：本文结合戴伊克新闻文本的话语图式，以体现文本重要事件信息的事件词所分布的句子为观测点，指出了突发事件文本由主线信息链和副线信息链构成。其中，明确提出主线信息链代表了文本的事件信息结构，由前核心事件链、核心事件链、次生事件链和再生事件链构成。副线信息链则是由“评价”部分、“背景”部分以及“情节”部分中的过细节信息以及事件词缺省的事件信息所在的句子等构成。同时，针对事件词有效性和区分度的问题，本文利用知网（HowNet）对事件词进行扩充，利用部分词语标记过滤副线信息链，实验结果证明方法具有可行性。

关键词：事件词；事件信息结构；话语图式；主线信息链；副线信息链

Analysis of Text's Event Information Structure by Event Word-driven

Zeng Qingqing, Yang Erhong
Institute of Applied Linguistics, Beijing Language and Culture University, Beijing 100083
E-mail: qing8612@sina.com; yerhong@126.com

Abstract: Event word plays an important role in event. The recognition of event word is the basic step in the field of information extraction. This paper tags all the event words from 680 event texts. Considering the annotation results and Van Dijk's Discourse Schema, it do some study on the information structure and the article defines that the sudden event discourse is composed of two elements, the main Information Chain and the second Information Chain. And the main Information Chain is just the text's event information structure including the former-core event information chain, the core event information chain, the secondary event information chain and the regeneration event information one. Also, we enlarge the number of event words in earthquake texts by using HowNet. Then, marked with some special words, we filter some secondary information.

Keywords: event word; event information structure; discourse schema; the main information chain; the second information chain

1　引言

随着互联网的广泛应用，新闻信息的利用需求不断提高，准确地从大量无序、杂乱、无结构的信息中提取用户感兴趣的事件信息已经成为信息抽取领域一个重要的研究课题。现有的事件抽取研究还主要局限在句子范围内，对整篇文本进行事件信息抽取的研究并不多。一个主题文本中往往包含了很多事件，其描述信息通常分散在整个文档中。

本文的研究立足于探索突发事件文本的事件信息结构，在篇章结构和信息结构之间建立联系，是语篇分析理论探索和自然语言处理相结合的尝试，试图为篇章理解研究做些探索性工作。

2　突发事件文本的篇章结构和事件信息结构

2.1　篇章结构

篇章结构是指通过语法手法、逻辑手法、修辞手法等将文章各个部分连接成为一个有机的整体。把握篇章结构，就能科学地揭示各段、各意群之间的关系，从而确立主题。在以抽取事件信息为目标的信息抽取任务中，信息抽取的主要对象指事件及其论元。由此，事件信息结构是由语

* 基金资助：国家社科基金项目“面向内容计算的文本信息标注研究”（06YY047）。

篇中包含事件及其相关论元在内的事件信息句子构成的结构链条。对于突发事件文本而言，文本的主题是事件本身，由此篇章结构和事件信息结构之间必然存在联系。2003 年，戴伊克（Van Dijk）在《作为话语的新闻》一书中具体概括了新闻文本的话语图式，如下图所示：

图 1　假设性新闻图式结构

结合上图戴伊克阐释的假设性新闻图式结构考虑可以发现，要了解文本发生的事件信息，需要阅读“主要事件”和“后果”组成的“情节”部分，可以忽略图式中的“背景”及“评价”信息。换言之，可以假定“情节”部分是突发事件的主体，也是信息抽取的主要部分。为此，本文选取了 200 篇火灾、200 地震、200 篇食物中毒和 80 篇恐怖袭击文本作为基础语料，人工标注文本中出现的事件词，得到每类文本的事件词集合[1]。

通过人工标注事件词的方式考察戴伊克（Van Dijk）的图式结构和本文探索的事件信息结构之间的关系，标注过程中发现：“情节”部分除却一些过于细节和事件词缺省的句子，基本上囊括了信息抽取所需的事件信息。为了区分戴伊克的图式结构，更加清晰地反映突发事件文本的篇章结构，并将这种篇章结构和文本的信息结构更好地对应起来，可以做这样的定义：

1. 主线信息链。一个独立语篇的主线信息链是指除去过细节信息以及事件词缺省的事件信息所在的句子的“情节”部分。从信息抽取角度来说，此信息链是由以事件词为显性标记的与文章报道核心事件相同或者相关的各类事件关联而成，是文本的中心部分，构成了篇章结构的主要部分，是读者进行篇章阅读和理解的最重要部分。

2. 副线信息链。副线信息链则是由“评价”部分、“背景”部分以及“情节”部分中的过细节信息以及事件词缺省的事件信息所在的句子等构成。从信息抽取的角度来说，副线信息链的信息不是信息抽取的关注对象。副线信息链的作用在于加深对新闻的认识和理解，深化新闻的主题。

用韦恩图表示主副线信息链和新闻图式结构成分的对应关系，表示如下：

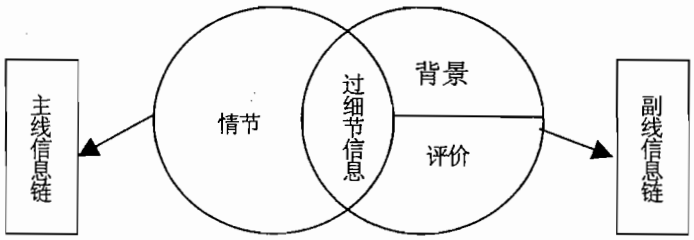

图 2　主副线信息链和新闻图式结构成分的对应关系

[1] 其中地震类文本事件词个数共 132 个，火灾类文本事件词个数 164 个，食物中毒类文本事件词个数 202 个，恐怖袭击类文本事件词个数 115 个。

2.2 事件信息结构

前文定义的主线信息链即为突发事件文本的信息结构，主线信息链上的句子能关联事件信息抽取所需的事件词和事件论元。通过对选择的所有语料的主线信息链进行意义和认知分析、考察，考察事件和事件之间的关系，可以发现主线信息链上的事件是分层次的。以获取的四个事件词集合为参照，主线信息链代表的事件信息结构是一个四级层次的事件框架体系，包括核心事件信息结构、前核心事件信息结构、次生事件信息结构以及再生事件信息结构。本文把这种结构也称为信息链。

（1）核心事件链。核心事件是事件信息结构中的重要构成成分，它是突发事件文本报道的焦点事件。标志核心事件发生的事件词即为核心事件词。包含核心事件词的事件小句是核心事件信息链。例如，地震文本的核心事件词集合如下：

Core Event Words of Earthquake =【地震、强震、大震、主震、微震、弱震、余震、震波、震感、有感、震央、震中、震级、震源、震深、震度、烈度】

（2）前核心事件信息链。本文所指的前核心事件指先于核心事件而发生的事件，大部分事件是造成核心事件发生的原因。前核心事件词在文中标识前核心事件的发生。包含前核心事件词的事件小句构成前核心事件信息链。例如，火灾类文本的前核心事件词集合如下：

Former-Core Event Words of Fire =【爆炸、点燃、短路、使用不当、操作不当、纵火、闪电、雷击、释放烟花、燃放烟花炮竹、取暖、泄露、拆除、熏制、焊接、超负荷、故障、争执】

（3）次生事件信息链。本文所指的次生事件是由核心事件直接造成的不可抗拒的事件，是事故造成的直接影响。次生事件词指次生事件的触发词。包含次生事件词的事件小句构成次生事件信息链。例如，食物中毒文本的次生事件词集合如下：

Secondary Event Words of Poisoned =【死亡、亡、死、伤、吃坏、上吐下泻、呕吐不止、口吐白沫、干呕、呕吐、作呕、反胃、吐泻、胀、肚子疼、拉肚子、急性肠胃炎、急性肠炎、急性胃炎、肠胃炎、肠胃不适、胃痛、肚痛、腹痛、胃胀、腹泻、腹泄、痢疾、身体不适、不舒服、发病、不适、难受、恶心、头痛、头疼、头晕、发烧、高烧、低热、发热、嗓子痛、休克、窒息、抽搐、昏迷、病重、危重、心悸、心慌、手抖、胸闷、虚脱、憔悴、浑身乏力、四肢无力、腿发软、站立不稳、手脚发麻、脸色潮红、面色苍白、脸色苍白、脸色发白、皮肤发红、过敏、胸闷气急、口唇紫绀、口唇发紫、心率加快、昏迷不醒、神志不清、意识不清、心有余悸、疼得难受】

（4）再生事件信息链。再生事件是指由核心事件造成的间接影响，主要是描述人在面对各种突发性的灾难时采取的各种应对措施。再生事件词指再生事件的触发词。包含再生事件词的事件小句组成再生事件信息链。例如，恐怖袭击文本的再生事件词集合如下：

Regeneration Events Words of Terrorism =【赶到、救援、增派、加强、抬进、送往、治疗、救治、搜捕、逮捕、搜出、被捕、追捕、击毙、打死、抓获、夺回、声称、宣布、声明、讲话、宣称、透露、影响、启动、警戒、呼吁、谴责、哀悼、清理、部署、疏散、戒严、封锁、撤离、拆除、停课、发现、统计、证实、调查、排查、戒备、盘查、勘察、勘查、立案】

3 事件词扩充和副线信息链过滤

3.1 事件词扩充

为验证从标注文本中标注得到的事件词集合对新的文本覆盖的有效性，本文做了一个简单的实验，将标注得到的事件词作为种子事件词，对新的测试语料文本进行事件词覆盖测试（先分词，但覆盖结果不考虑分词错误）。以地震文本为例，重新选择 50 篇新的文本。覆盖结果验证了地震文本的种子事件词不能完全覆盖新文本，即新文本中出现了新的事件词。因而，所获得的事件词

对同类事件新闻报道文本信息表达的有效性不够。本文的解决方法是不增加标注量，而是利用现有常识知识库知网（HowNet）对已有的种子事件词进行扩充。在知网（HowNet）已有的 Concept Relevant Calculator 执行程序中，可以利用该执行程序检索出某一个词的相关词。

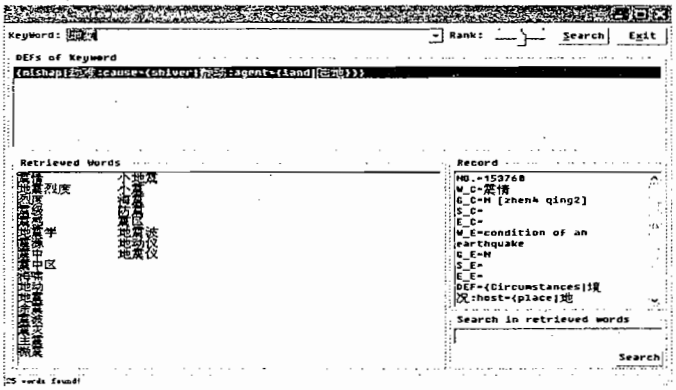

图 3 “地震”在知网（HowNet）中的 Rank II 相关词

结合知网（HowNet），本文扩充方法如下：利用 perl 脚本语言从知网（HowNet）中获得种子词的 Rank II 相关词；利用中科院自动化所分词软件（2006 版）对检索结果进行词性标记，利用部分词性标记（如 LOC、ORG）过滤不可能成为事件词的相关检索词；最后人工校剩余结果，得到最终扩充词集合。以地震文本为例，核心事件词经扩充后由原来的 17 个扩展为 21 个[1]；地震文本次生事件词由原来的 64 个扩展为 1146 个；地震文本再生事件词由原来的 51 个，扩展为 548 个。

在对四类文本的其他事件词进行扩充时，扩充原则一样，但是四类文本的前核心、再生、次生事件词之间有很多交集词语，因此对每类事件词类别可以区别对待：核心事件词需要分门别类，具体对地震、火灾、食物中毒、恐怖袭击文本的核心种子事件词分别进行扩充，得到扩充后事件词集合；其次，可以将标注得到的四类文本的前核心、再生、次生事件词作为一个整体进行词语扩充，而不再需要分别重复扩充每类文本的所有事件词。

3.2 副线信息链过滤

人工标注过程中，已经发现副线信息链中很多句子包含事件词，诸如背景信息、评价信息等等。这种现象说明事件词对信息结构的区分度并不够。为了提高事件词对信息结构的区分度，需要对各类文本的副线信息链信息进行过滤。所谓信息过滤是指过滤文本中影响事件抽取的干扰信息。要过滤副线信息链信息，将它们提取出来放入暂不处理的集合当中，特征的选取和发现尤为关键，这将涉及到究竟什么样的特征能够有代表性的区分开主线信息链和副线信息链的句子。

一般来说，过细节信息属于客观信息的一部分，但是因为其过于琐碎，往往句子当中不会包含有标注和扩充得到的事件词，所以对于过细节信息可以暂不考虑。例如：

（1）罗周忠因外出不在家，逃过一劫，罗还有一个女儿在外地读书。

（2）两个穿白衬衣黑西裤的男子一见记者，就警惕地问：“你们是干什么的？”当摄影记者拿出相机时，立刻上来三四个男子将他堵在角落，有人扬言要“多找几个人来打你”。

另外，有一些背景信息也不包含事件词，不会对事件抽取造成干扰，例如：

（1）呼图壁县位于新疆中北部，距离新疆首府乌鲁木齐约六十公里。

（2）巴达赫尚省是阿富汗最偏远的地区，交通不便、通信落后、人口密度很低。

[1] 地震核心事件词扩充的效果并不是十分明显，这是因为在人工标注地震类文本的时候，就已经标出了大部分代表地震事件的事件词了。

过滤的重点是包含事件词的评价信息和背景信息。本文主要采取词语的显性标记作为过滤手段。例如，在标注过程中发现，地震文本的背景信息有比较明显的特征。对 200 篇地震文本考察，发现很多背景信息表达方式如下：

（1）日本地震频发，每年发生有感地震 1000 多次，是世界上地震最频繁的国家之一。

（2）墨西哥处于环太平洋地震带东部，属地震多发国家。

（3）地处太平洋板块和加勒比板块交界处的尼加拉瓜境内地壳运动频繁，历史上曾多次发生地震。

（4）去年 8 月，秘鲁发生里氏 8 级地震，至少造成 500 人死亡，4 万座房屋被毁。

从这些包含知识、历史、环境以及以前事件在内的背景信息中，可以找到诸如“(频繁)|(频发)|(多发国家)|(多发区)|(多发带)|(多发地带)|(强地震带)|(最易发生)|(经常发生)|(活跃)|(曾发生)|(曾多次发生)|(曾遭遇)|(发生过)|(上次发生)|(上一次发生)|(去年)”这样的词语显性标记。在选取的 200 篇地震文本中，人工标记有 59 个句子是背景信息。通过 perl 程序将以上出现的词语作为抽取显性标记，能够抽出 45 个句子，抽全率为 76.27%。

又如，对于一些评价信息，也可以采用词语作为标记，例如：

（1）分析人士认为，不管调查结果如何，巴基斯坦的国际形象因这次袭击事件而再次遭受严重影响，使外界对巴基斯坦的安全形势感到进一步担忧。

（2）警方初步判断是泰南武装分子制造了这起恶性恐怖袭击事件。

（3）估计在未来 24 小时内，景泰原震区发生更大级别地震的可能性不大。

（4）伊朗驻联合国官员的一系列可疑行为已引发了纽约警局官员有关伊朗特工可能主使发动恐怖袭击的担心。

3.3 实验

为验证扩充效果，了解过滤副线信息的用途，设计如下实验：依旧以原有的 200 篇地震文本作为训练语料，50 篇新的地震文本作为测试语料，分别做开放和封闭测试。步骤：1. 在进行覆盖之前，先尽可能利用规则过滤副线信息链；2. 然后利用扩充后的所有事件词对文本进行覆盖识别。

表 1　地震类文本事件词扩充前封闭及开放测试实验（且未过滤副线信息）

地震类文本事件词识别	Precision	Recall	F-Score
封闭测试	89.68%		
开放测试	90.02%	97.60%	93.66%

表 2　地震类文本事件词扩充后封闭及开放测试实验（且过滤副线信息）

地震类文本事件词识别	Precision	Recall	F-Score
封闭测试	95.57%		
开放测试	92.24%	99.15%	95.57%

在表 1 中，开放测试准确率高于封闭测试，是因为选择的语料量比较少，则副线信息链会相对少，错误率也会低一些。从以上两个表的实验数据可以看出，通过过滤和事件词扩充二个步骤，一方面减少了错误识别结果，提高了正确识别率，另一方面，因为扩充后的事件词集扩大，因而使得更多的事件词能够被机器识别出来，召回率也得到了提高。表中 F-值提高效果比较明显。

4 结语和下一步工作

本文结合戴伊克新闻文本的话语图式，以体现文本重要事件信息的事件词所分布的句子为观测点，指出了突发事件文本由主线信息链和副线信息链构成。其中，明确提出主线信息链代表了

文本的事件信息结构，由前核心事件链、核心事件链、次生事件链和再生事件链构成。副线信息链则是由“评价”部分、“背景”部分以及“情节”部分中的过细节信息以及事件词缺省的事件信息所在的句子等构成。同时，针对事件词有效性和区分度的问题，本文利用知网（HowNet）对事件词进行扩充，利用部分词语标记过滤副线信息链，实验结果证明方法具有可行性。

但是本文的研究只局限于地震、火灾、食物中毒、恐怖袭击四类文本，对任何其他新的事件类型是否适用还有待考察。新闻报道的角度不同，风格不一，所作的工作远不是本文所能涵盖的；突发事件文本副线信息链的过滤规则过少。下一步工作将继续对更多文本进行考察。

参 考 文 献

[1] Ralph Grishman. Information Extraction: Techniques and Challenges [M]. Information Extraction, ed. Maria Teresa Pazienza, Spring Notes in Artificial Intelligences, Spring-Vealag. 1997.

[2] ACE. ACE Chinese Annotation Guidelines for Entities (Version 5.5) [EB/OL].

[3] http://www.ldc.upenn.edu/Projects/ACE/docs/Chinese-Entities-Guidelines_v5.5.pdf. 2005a.

[4] ACE. ACE Chinese Annotation Guidelines for Relations (Version 5.5.1) [EB/OL].

[5] http://www.ldc.upenn.edu/Projects/ACE/docs/Chinese-Relations-Guidelines_v5.5.1.pdf.2005b.

[6] ACE. ACE Chinese Annotation Guidelines for Events [EB/OL].

[7] http://www.ldc.upenn.edu/Projects/ACE/docs/Chinese-Events-Guidelines_v5.5.1.pdf. 2005c.

[8] 杨尔弘. 突发事件信息提取研究[D]. 北京语言大学, 2005.

[9] 袁毓林. 信息抽取的语义知识资源研究[J]. 中文信息学报, 2002.

[10] 姜吉发. 自由文本的信息抽取模式获取的研究[D]. 中国科学院研究生院（计算技术研究所）, 2004.

[11] 董振东, 董强. 知网（HowNet）. http://www.keenage.com.

[12] 赵妍妍, 秦兵, 车万翔, 刘挺. 中文事件抽取技术研究[J]. 中文信息学报, 2008.

[13] [荷]Van Dijk（著），曾庆香（译）. 作为话语的新闻[M]. 华夏出版社, 2003.

[14] 钱敏汝. 戴伊克的话语宏观结构论（上）[J]. 国外语言学, 1988.

[15] 钱敏汝. 戴伊克的话语宏观结构论（下）[J]. 国外语言学, 1988.