

基于依存分析的事件识别

付剑锋¹ 刘宗田¹ 付雪峰² 周 文¹ 仲兆满¹

(上海大学计算机工程与科学学院 上海 200027)¹ (南昌工程学院计算机科学与技术系 南昌 330099)²

摘 要 事件抽取是信息抽取的重要组成部分,事件识别是事件抽取的基础,事件识别的效果直接影响了事件抽取的结果。基于机器学习的方法识别事件需要从词汇中发掘更多的特征。针对当前事件识别方法中存在的不足,提出了一种基于依存分析的事件识别方法。用依存分析发掘触发词与其它词之间的句法关系,以此为特征在 SVM 分类器上对事件进行分类,最终实现事件识别。实验表明,基于依存分析的事件识别优于传统的事件识别方法,而融合多特征的事件识别 F 值可提高到 69.3%。

关键词 事件识别,依存分析,支持向量机

中图法分类号 TP391 文献标识码 A

Dependency Parsing Based Event Recognition

FU Jian-feng¹ LIU Zong-tian¹ FU Xue-feng² ZHOU Wen¹ ZHONG Zhao-man¹

(School of Computer Engineering & Science, Shanghai University, Shanghai 200072, China)¹

(Department of Computer Science, Nanchang Institute of Technology, Nanchang 330099, China)²

Abstract Event Extraction is an important part of information extraction. As the basis of Event Extraction, Event Recognition directly affects the results of Event Extraction. Machine learning based Event Recognition needs to find more features in words. For the deficiency of present Event Recognition method, this paper presented a novel method of Dependency Parsing based Event Recognition (DPER). Dependency parsing was used to find the syntactic relation among triggers and other words. As one of features, this relation was used to event classification on SVM and then to event recognition. The experiments show DPER has better performance than traditional method, and Event Recognition integrating multi-features improves F -measure to 69.3%.

Keywords Event recognition, Dependency parsing, SVM

1 引言

信息抽取是当前自然语言处理中的一个重要分支,事件抽取是信息抽取的重要组成部分。事件抽取就是从非结构化文档中抽取出用户感兴趣的事件,同时用结构化的形式描述,供用户查询及进一步分析。事件抽取在自动文摘^[1,2]、问题回答系统^[3]等方面有着广泛的应用。事件抽取主要包含两个步骤,第一步是事件的识别;第二步是对识别出来的事件进行分析,进而抽取其中的事件要素,这些要素包括事件发生的时间、地点、事件的参与者等。其中事件识别是事件抽取的基础,事件识别的效果直接影响了事件抽取的结果。

依存分析可以发现词语之间的句法和语义关系,在问题回答、机器翻译、信息检索等方面有着广泛的应用^[4]。本文首先综述了事件抽取的相关工作,分析了当前事件抽取的主流方法,根据现有方法中存在的不足,提出了一种基于依存关系的事件识别(DPER)方法。实验表明,此方法可以有效提高

事件识别的 F 值。

2 相关工作

美国国防高级研究计划委员会(Defense Advanced Research Projects Agency, DARPA)主办的消息理解会议(Message Understanding Conference, MUC)和话题识别与跟踪(Topic detection and tracking, TDT),以及由美国国家标准技术研究所(National Institute of Standards and Technology, NIST)组织的自动内容抽取(Automatic Content Extraction, ACE)评测会议都将事件的识别与抽取作为其评测任务之一。这些会议在推动信息抽取技术发展的同时,也将事件抽取的研究变成信息抽取中的一个热点。

目前国内外研究事件抽取的方法主要可以分为两大类:基于规则的方法和基于统计学习的方法。当然也有结合两者,使用混合算法的。文献[5]针对生物医学领域,根据该领域文献语法特征构造了一系列规则,封装成抽取器(Extrac-

到稿日期:2008-12-15 返修日期:2009-02-25 本文受国家自然科学基金(60575035),上海高校选拔培养优秀青年教师科研专项基金(shu-07027)和上海市重点学科建设项目(J50103)资助。

付剑锋(1978-),男,博士研究生,主要研究方向为自然语言处理,E-mail:feng93017@tom.com;刘宗田(1946-),男,教授,博士生导师,主要研究方向为智能信息处理、软件工程;付雪峰(1978-),男,硕士,讲师,主要研究方向为信息检索;周 文(1979-),女,博士,讲师,主要研究方向为概念格、事件本体;仲兆满(1977-),男,博士研究生,讲师,主要研究方向为自然语言处理。

tor),抽取其中的事件。文献[6]对在线新闻中的气象事件构造规则,抽取关于气象方面的事件。文献[7]用 MegaM 作为二元分类器和 TiMBL 作为多元分类器两种机器学习方法实现了事件抽取,在 ACE 英文语料上取得了不错的效果。在中文的事件信息抽取方面,文献[8]利用手工确定的句型模板构造了抽取规则,用于从处理后的文本中抽取事件信息填充句型模板中的槽。文献[9]通过语句聚类的方法获得事件的信息结构(事件模板),以抽取相应事件。文献[10]采用机器学习的方法改进了文献[7]中训练集正反例不平衡以及数据稀疏的不足,在 ACE 中文语料上取得了较好的效果。

基于规则的方法在特定领域内可以取得比较好的效果,但是规则的可移植性差,从一个领域移植到另一个领域时,需要重新构建规则。而且规则的制定费时费力,需要领域专家的指导。虽然引入机器学习的方法可以从一定程度上加速规则的制定,但是不同规则之间造成的冲突也是一个棘手的问题。采用基于统计的机器学习的方法,与特定的领域无关,可以减少人工干预的过程,也不太需要领域专家的指导,并且可移植性较好。机器学习的方法需要一定的语料作为学习用例和测试用例。不过,随着语料库建设的不断发展以及互联网上各种文本资源不断丰富,语料的获取不再是束缚机器学习的瓶颈。因此,机器学习的方法逐渐成为当前事件抽取的主流技术。

采用机器学习的方法识别事件,就是借鉴文本分类的思想,将事件的识别转化成为分类问题。与文本分类不同的是,文本分类通常需要选择特征,对特征降维。而事件常常以句子为单位(也可能是一个句子包含多个事件),所以事件一般包含较少的词汇信息,这给事件分类带来了一定的困难,因此需要从句子中充分发掘词汇各种潜在的特征来提高分类性能。针对这种情况,本文在传统事件识别利用词汇、词性信息构造特征向量的基础上,引入句子的依存关系来构造特征向量。采用 SVM(支持向量机)机器学习算法实现事件识别,通过实验验证了引入依存关系后的事件识别性能。

3 依存分析

依存的概念来自句法结构中不对称的二元词汇关系的思想,1959年由法国语言学家 L. Tesnière 首先提出。他认为,二元词语关系存在两个词语之间,依存语法由不对称的二元词语关系构成,这种二元词语关系被称为依存关系。依存关系包含两个成分:核心词(head)和依存词(dependent),依存关系反映出核心词和依存词之间语义上的依赖关系。依存关系是有方向的,表现为一个词支配另一个词,这种支配和被支配的关系和语义相关联,并且不受距离的约束。语义依存分析的目的就是发现词语之间的这种语义联系。对例1进行依存分析后,结果如图1所示。

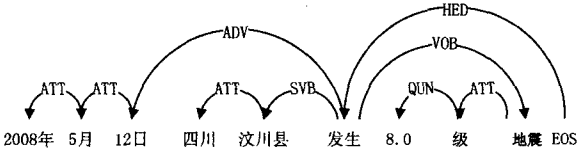


图1 依存分析示例

例1 2008年5月12日,四川汶川县发生8.0级地震。
图1中 HED 表示“发生”是句子的核心动词,ATT 表示

定中关系,ADV 表示状中结构,SVB 表示主谓关系,VOB 表示动宾关系,QUN 表示数量关系。两个词之间的依存关系由一个有向弧表示,弧的发起端的词(依存词)以某种关系依存于弧的指向端的词(核心词)。可以看出,依存关系是建立在分词之上的,比词性标注更为深层次地反映出句子语法结构的一种关系。本文采用了哈工大 LTP 平台实现依存分析,LTP 对句子处理后的输出是一个以词为单位的三元组链表,每个元组都可以表示为 (POS, DR, PID),其中 POS 表示词性,DR 表示依存关系,PID 表示父节点编号。例1经过 LTP 处理后,输出的结果如表1所列。

表1 LTP 输出结果

词	词性	依存关系	父节点编号
2008 年	nt	ATT	1
5 月	nt	ATT	2
12 日	nt	ADV	6
四川	ns	ATT	5
汶川县	ns	SBV	6
发生	v	HED	-1
8.0	m	QUN	8
级	q	ATT	9
地震	n	VOB	6

4 事件识别

“事件”这一概念经常出现在哲学、认知科学、语言学、人工智能等领域的文献中,然而人们对事件一直都没有一个确切的定义。不同学科之间由于关注的内容和研究的重点不同,定义各不相同。即使在同一个领域(比如人工智能领域),也存在不同的观点。ACE 中定义事件为包含参与者参与的特殊的事情,事件通常可以描述为一种状态的改变。本文的事件识别参考了 ACE 的评测内容,给出相关定义如下。

定义1(事件,Event) 在某个特定的时间和地点下发生、由若干角色参与、表现出若干动作特征的一件事情。其中时间、地点、事件参与的对象称为事件要素。

这里定义的事件是指现实世界所发生的事情。现实世界的事件和语言表达中的事件有所不同,现实世界的事件和事件要素都是客观存在的,而语言中的事件根据表达的需要,其要素可能缺省或者缺失。为了方便,将语言中的事件也称之为事件。

定义2(事件触发词,Trigger)^[11] 可以用来清晰地表示所发生的事情的词。一般情况下,触发词是句子中的主要动词(也可能是名词),触发词直接描述了事件。

例2 截至10日12时,四川汶川地震已造成69197人死亡。

例3 5月12日,受强烈地震影响,兰州市区感受到明显震感。

在例2中,包含了地震和死亡两个事件,斜体部分标识出了地震事件和死亡事件的触发词,其中地震为名词,死亡为动词。事件触发词并不唯一,在例3中,震感表示兰州市区发生了地震事件。

定义3(事件识别,Event Recognition) 从包含触发词的句子(文本)中找出现实世界发生的事件。

例4 地震是一种正常的自然现象。

在例2和例3中,触发词都可以清晰地表示其所发生的事情。但是,包含触发词并不意味着事件的发生。同一个触

发词在不同的上下文语境中,所代表的意义是不同的。例 4 中的地震就不表示发生了地震事件,它只是对地震的一种描述。因此,孤立地考虑触发词和事件之间的关系,不可能正确地识别出事件。必须结合上下文特征,把触发词、与触发词相近的词以及这些词的词性、位置信息、依存关系等作为一个整体来考虑,根据这些特征来识别事件。把事件识别的问题当作事件分类的问题,实际上是借鉴文本分类的思想,判断包含触发词的句子是否能划归于某一类事件。所以,从分类的角度可以定义事件识别为:把给定的某个事件映射到已知的 K 个事件类型中的某一个,一个事件只能属于一个事件类别。事件识别过程如图 2 所示。

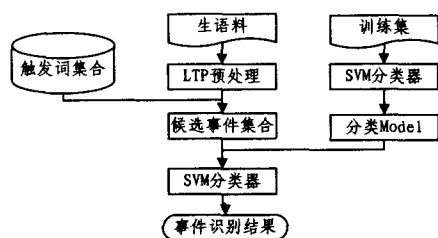


图 2 事件识别过程

本文选取的与触发词相关的特征如下:

- 1) 触发词以及触发词的词性;
- 2) 触发词左侧 8 个词及其词性;
- 3) 触发词右侧 9 个词及其词性;
- 4) 上述词之间的依存关系。

5 实验结果及分析

为了评价 DPER 的性能,采用了 SVM 作为分类器,用 Java 语言实现了 DPER。SVM 分类器采用 libSVM,从 <http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html> 下载。通过实验与传统的没有采用依存关系作为特征的事件识别效果进行了比对。实验语料来自于 Web 上收集的各类事件的报道,共包括地震、台风和海啸等 8 类事件的 416 篇文档。其中 300 篇作为训练语料、116 篇作为测试语料,并建立了相关事件的触发词库。采用准确率(precision)、召回率(recall)和 F 值(F -measure)3 个指标来评价实验结果,实验结果如表 3 所列。

表 2 事件识别实验结果

Features	precision	Recall	F-measure
Word	58.4%	56.2%	57.3%
Word+POS	65.6%	60.3%	62.8%
Word+DR	67.7%	63.5%	65.5%
Word+POS+DR	71.6%	67.2%	69.3%

从实验结果可以看出,仅用词(Word)作为特征,只能得到 57.3%的 F 值。加入词性标注之后(Word+POS),将 F 值提高到 62.8%。词加上依存关系(Word+DR),可以得到 65.5%的 F 值。综合 3 个特征(Word+POS+DR),将 F 值提高到了 69.3%。分析实验结果,可以得出如下结论。

1) 选择依存分析(Word+DR)作为事件识别的特征,分类效果好于用词性(Word+POS)作为特征。这是因为依存关系比词性能更深层次地反映句子的语法结构,因此依存关系具有比词性更好的分类性能。

2) 以单纯使用词的分类结果为基准,在事件识别时,选择的特征越多,分类的性能越高,事件识别的效果也越好。事件识别由于其本身的特殊性,一个事件包含的词汇非常有限,因此需要在有限的词汇中发掘更多的特征来区分事件,特别是发掘词汇的语法和语义特征。

结束语 本文针对当前事件识别中存在的不足,提出了一种基于依存分析的事件识别方法。实验表明,选择依存关系作为事件的特征,可以有效提高事件识别的 F 值。不过,从实验的最后结果可以看出,目前事件识别的效果还有很大的提升空间。基于机器学习的事件识别方法综合了很多自然语言处理技术中的基础工作,比如说分词和依存分析。分词和依存分析的结果,对事件识别的效果有着直接的影响。因此,进一步提高分词和依存分析的正确率,可以从一定程度上提高事件识别的效果。进一步发掘事件中各个词汇之间的语义关系以及融合多种特征识别事件,是下一步的研究方向。

参考文献

- [1] Daniel N, Radev D, Allison T. Sub-event based multi-document summarization[C]// Association for Computational Linguistics Morristown, NJ, USA, 2003: 9-16
- [2] Filatova E, Hatzivassiloglou V. Event-based Extractive Summarization[C]// Association for Computational Linguistics, 2004: 104-111
- [3] Yang H, Chua T S, Wang S, et al. Structured use of external knowledge for event-based open domain question answering [M]. New York, NY, USA: ACM Press, 2003: 33-40
- [4] Nivre J, Scholz M. Deterministic dependency parsing of English text[C]// Association for Computational Linguistics Morristown, NJ, USA, 2004: 64-70
- [5] Yakushiji A, Tateisi Y, Miyao Y, et al. Event extraction from bio-medical papers using a full parser. 2001: 408-419
- [6] Lee C S, Chen Y J, Jian Z W. Ontology-based fuzzy event extraction agent for Chinese e-news summarization[J]. Expert Systems with Applications, 2003, 25(3): 431-447
- [7] Ahn D. The stages of event extraction[C]// Proceedings of the COLING-ACL 2006 Workshop on Annotating and Reasoning About Time and Events. 2006: 1-8
- [8] 吴平博, 陈群秀, 马亮. 基于事件框架的事件相关文档的智能检索研究[J]. 中文信息学报, 2003, 17(06): 25-30
- [9] 杨尔弘. 突发事件信息提取研究[D]. 北京: 北京语言大学, 2005
- [10] 赵妍妍, 秦兵, 车万翔, 等. 中文事件抽取技术研究[J]. 中文信息学报, 2008, 22(1): 3-8
- [11] Consortium L D. ACE(Automatic Content Extraction) English Annotation Guidelines for Events. 2005