

中文事件抽取中事件类别的自动识别

赵妍妍, 王啸吟, 秦兵, 车万翔, 刘挺

(哈尔滨工业大学计算机学院信息检索研究室, 哈尔滨 150001)

摘要: 事件抽取是目前信息抽取研究领域的一个新的重要的研究课题。本文结合美国国家标准技术研究院(NIST)组织的自动内容抽取(ACE, Automatic Content Extraction)评测中的事件抽取任务的要求, 在 ACE2005 的训练数据上进行事件抽取中事件类别识别的实验。实验中采用《同义词词林(扩展版)》扩展从训练语料中提取出的触发词, 构建触发词表, 并结合两种机器学习方法——最大熵(ME, Maximum Entropy)和支持向量机(SVM, Support Vector Machine), 抽取合适的特征, 使得事件类别识别的 F-Score 分别达到了 69.2%和 69.9%。

关键词: 事件抽取; ACE 评测; 最大熵; 支持向量机

Automatic Event Type Extraction in Chinese Event Extraction

Yanyan Zhao, Xiaoyin Wang, Bing Qin, Wanxiang Che, Ting Liu

(Information Retrieval Laboratory, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract: Event Extraction is a new research point in the area of Information Extraction. In this paper, we carried out a series of experiments on event detection and classification based on the standard and training corpus of ACE05 (Automatic Content Extraction 2005), which is a research task organized by NIST (National Institute of Standards and Technology). In the experiment, a table of trigger words was constructed by extracting words from training corpus directly, and was extended by *Tongyici Cilin(extended)*. Further more, two approaches of machine learning——ME(Maximum Entropy) and SVM(Support Vector Machine) were used to acquire a higher precision. The F-score of our result reaches 69.2% with ME and 69.9 % with SVM respectively.

Keywords: Event Extraction; ACE tasks; Maximum Entropy; Support Vector Machine

1 引言

事件抽取是目前信息抽取研究领域的一个新的研究点。信息抽取的主要目的是将无结构的文本转化为结构化或半结构化的信息, 并以数据库的形式存储, 供用户查询以及进一步分析利用。信息抽取系统的主要功能是从文本中抽取特定的事实信息, 我们称之为实体(Entity), 例如: 时间(TIME)、组织机构(ORG)、人物(PER)等等。多个实体在一定的条件下可组成各种不同类型及子类型(Type/Subtype)的事件。确定事件的类型、子类型以及所包含的实体在事件中扮演的角色(Role), 我们称之为事件抽取^[1]。事件抽取用于篇章级的抽取中,

基金资助: 国家自然科学基金, 资助号: 60435020, 60575042, 60503072; 腾讯基金项目

作者简介: 赵妍妍(1983-), 女, 山东人, 硕士研究生 email: zyy@ir-lab.org

能够比较好的概括出一篇或几篇文章的主要事件，因此可以应用在多文档文摘^[2]，自动问答^[3]等研究领域里。

事件抽取任务可分为两步，一是确定事件的类别以及子类别，二是确定当前类别的事件所包含的实体参与者及其角色。本文主要结合美国国家标准技术研究院(NIST)组织的自动内容抽取(ACE, Automatic Content Extraction)评测中的事件抽取任务的要求，在ACE2005的训练数据上进行事件抽取中事件类别识别的实验。

与实体和实体关系的抽取类似^[4]，事件抽取中事件的类型和子类型也是预先定义的。比如：公司合并事件(Business/Merge-Org)、袭击事件(Conflict/Attack)等等。映射到文本中，如：“胡锦涛和布什举行了会谈。”，这就是一个事件的实例，本文的任务就是要判断事件的类型/子类型。其中，“会谈”这个词能够很好的概括出该事件的中心思想，我们把这类词称为触发词(trigger)，“胡锦涛”，“布什”是填充这个事件的要素(argument)实体，扮演一定的角色(role)，通过判断此事件的类型/子类型为Contact/Meet事件。可见，事件的触发词对于事件类别的确定具有重要的作用。

目前，不少组织和个人也在从事事件抽取的相关工作，如：结合句法分析和手工模版的方法完成足球事件的抽取系统^[5]；基于限定域Ontology的气象事件抽取系统^[6]。一般来说，主要使用两种方法来实现对事件类别的识别。一是基于规则的方法^[7]，二是基于模板的方法。前者需要专家构筑大规模的知识库，这不但需要有专业技能的专家，也需要付出大量劳动；后者虽然解决了浪费劳力的缺点，但却同样不能跨领域使用。为了克服这两方面缺点，本文提出了结合触发词表使用机器学习的方法，然后，构造分类器。该方法不需要有专业技能的专家书写知识库，只需要有一定专业知识的人对事件的类型做出判断即可。除此之外，从训练语料中提取的触发词较少也是一个很棘手的问题，为了提高识别率，针对具体问题，我们采用了哈工大信息检索研究室的《同义词词林(扩展版)》来扩展触发词，从而也扩大了我们的正例和反例的数量，使得分类比较准确。

本文组织如下：第二部分简要介绍事件类别识别的方案设计；第三部分详细介绍基于扩展的触发词表和机器学习相结合的方法；第四部分给出实验结果和分析；第五部分给出结论和展望。

2 方案设计

本文主要实现事件类别识别系统，该系统读入一篇生文本和该文本的实体抽取结果，按照 ACE05 确定的事件分类的标准，从中识别出属于这 8 类的事件。系统主要由预处理模块、词表生成模块和分类模块组成，其中预处理模块读入一篇生文本，通过分句、分词处理和实体抽取，将其转化为一篇结构化的文本，然后查询触发词表过滤出包含触发词的句子，得到候选事件集合；词表生成模块通过读入大量的训练语料找出某些特定词汇与事件之间的联系，并按照一定的规则将这些词提取出来，形成触发词表；而分类模块从候选事件中抽取特征，形成特征向量，并调用机器学习软件包实现对候选事件的分类，区分出真正的事件。

系统的结构图如图 1 所示：

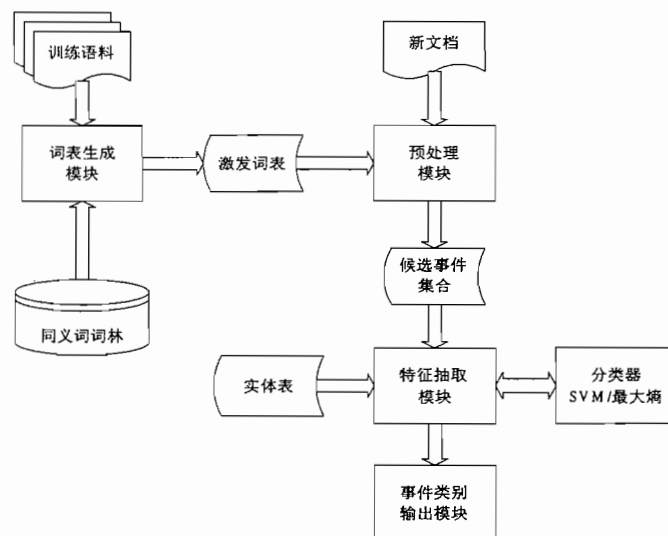


图 1 系统设计方案

Figure1 System Design

3 扩展触发词表和机器学习相结合的事件类别识别

在实验阶段中，我们主要对词表生成模块和分类模块进行了研究，分别尝试了基于扩展触发词表、基于机器学习识别、基于扩展的触发词表和机器学习相结合三种方法，以下是三种方法详述。

3.1 基于扩展触发词表的事件类别识别

事件中的触发词对识别事件的类型具有重要的作用，所以首先我们利用触发词来识别事件的类型。用训练数据中所有事件的触发词（训练语料中已标注）构造一张触发词表，每一项标明触发词及其所在的事件类别和子类别。如果待测试语料事件含有表中的触发词，则标为其所在的类别和子类别。例如：胡锦涛访问了美国。查询触发词表，发现“访问”在表中，则是事件，记录下其类别与子类别，否则，不是事件。在触发词表较大的情况下，这种方法的召回率较高，但是精确率不够理想，因为有很多词存在一词多义现象，如：“伤害”，据 ACE 标准规定，只有在伤害到肉体的时候才为 Life/Injure 的触发词，这时“伤害”所在的候选事件才是事件，否则不是事件。

为解决以上问题，滤去干扰的触发词，本文提出了一种类类似于 $TF*IDF$ 的方法，给每个触发词都计算一个分数 ($Score$)，分数越高，说明该触发词与其所对应类别的相关度越高。计算公式如下：

$$Score = ETF * EIDF \quad (1)$$

$$EIDF = \log_2 \left(\frac{N}{EDF} \right) \quad (2)$$

其中 ETF 为该触发词 tw 在所有训练语料中所产生的某类事件的总个数 / 这类事件的总个数，反映的是触发词 tw 对整个事件的贡献程度；公式(2)反映了该触发词 tw 在训练语料中出现的频繁程度，其中 N 为训练语料的句子总数， EDF 为含有该触发词的句子总数（事件抽取以句子为单位）。

这时， $ETF*EIDF$ 就可以反映出该触发词在整个训练语料中对类别的贡献大小， $Score$ 值越大，则说明词义越单一，代表某类事件的能力越强。这样可以根据 $ETF*EIDF$ 规定出一个阈值，过滤小于阈值的干扰词汇。实验表明：这一方法有效提高了精确率，解决了部分一词多义现象带来的干扰，但牺牲了部分的召回率。由于语料规模较小，好多新的触发词无法召回，所以本文使用《同义词词林（扩展版）》来扩充触发词表，提高召回率，并同时解决了语料小，触发词表规模有限的缺点。

《同义词词林（扩展版）》是在《同义词词林》的基础上，整理扩展而成的同义词机读词典，我们可以通过获得触发词的同义词扩展出新的触发词，如“出生”的同义词为：

诞生 出生 降生 生 落地 坠地 出世

这些词都是表示出生这一含义的词，并且可能触发一个“出生”类事件。

扩展的具体方案如下：对于触发词表中的每一个词，在同义词词林中查出它的全部义项，若某义项中的所有同义词中有 n 个或 n 个以上都在触发词表中，且这几个词的类别相同，则将这个义项中所有词汇全部扩展到触发词表中，并赋予前述相同的类别， n 称为扩展阈，在实验中取值为 2 和 3。

实验表明，这种方法的特点是召回率非常高，而准确率相对要低许多，F-Score 值不高。原因在于单独使用触发词不足以表达类别信息，因此需要加入更多与类别信息有关的特征帮助分类。

3.2 基于机器学习的事件类别识别

由于一词多义现象普遍存在，使用滤去干扰词汇的扩展触发词表的方法依然不能解决大部分问题。把所有含有触发词表中词的事件称为候选事件，因此候选事件有可能是事件，也有可能不是事件。仍以“伤害”为例，有两个词义，一是肉体上的伤害，一是精神上的伤害。直接使用词表很明显会造成事件的多标，考察“伤害”这个词周围的一些信息，即使用机器学习的方法选取恰当的特征，准确的标出一个含有“伤害”的候选事件是否是所规定的事件。

根据 ACE 事件抽取的评价标准，一个带有触发词的句子是否是所规定类型的事件，可看成是一个二值分类问题，即含有表中触发词的候选事件需要通过恰当的特征来确定。经过分析，发现触发词周围的实体的类型、词的词性，以及触发词的子类型都对事件的正确分类起到举足轻重的作用。

对一个包含触发词（《同义词词林（扩展版）》扩展后的触发词表中的词）的上下文构造的特征向量如图 2 所示：

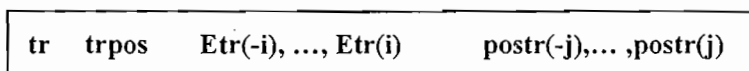


图 2 特征向量的构造

Figure 2 Construction of the feature vector

其中, **tr** 和 **trpos** 分别表示触发词本身和触发词的词性; **Etr(-i),...,Etr(i)**为触发词前后各 *i* 个实体的类型; **postr(-j),...,postr(j)** 为触发词前后各 *j* 个词的词性。

下面以 $i=2, j=3$ 为例:

这名高中生是在前天晚上在东京的涩谷娱乐区打伤8人后被捕的。

抽取结果为:

1:被捕 2:/v 3:GPE/Population-Center 4:PER/Group 5:N/A 6:N/A 7:/m 8:/n 9:/nd 10:/u 11:/wp 12:N/A

句子中加粗的“被捕”为触发词,下划线的“涩谷娱乐区”和“8人”为上下文中的实体,其类型分别为 *GPE/Population-Center* 和 *PER/Group*,触发词后面因为没有实体,所以用 *N/A* 表示,实体和事件类型的定义见参考文献[1]。

经实验证明,当实体的抽取窗口 *i* 调整为 1,而词性的抽取窗口 *j* 调整为 4 的时候实验结果最好。

本文主要采用了两种机器学习方法,最大熵(ME)和 SVM。ME 在自然语言处理研究中应用非常广泛,在机器翻译、句法分析及命名实体抽取等问题中收到了良好的效果^[8]。新加坡的 Hai Leong Chieu 和 Hwee Tou Ng 首次在事件抽取中引入该分类器,用于特定类别事件的要素(Argument)抽取^[9]。该模型简单,不需要特征独立假设,速度快的优越性非常适合于本文的事件类别识别;而 SVM 在解决小样本、非线性及高维模式识别问题中很有优势,非常适合事件抽取语料较少、情况复杂的特点。

由于机器学习方法的引入,使得事件类别识别的精确率有很大幅度的升高,但是召回率有所降低,因为机器学习是一种统计学习方法,不可能照顾到每一个实例,因此影响了召回率。

3.3 基于扩展的触发词表和机器学习相结合的方法

实验表明,基于扩展的触发词表的方法召回率非常高,而准确率比较低,而基于机器学习的类别识别方法虽然效果较前一种理想,但召回率却不高。因此考虑将这两种方法相结合,扬长避短。

方法设计如下:

- 1、将扩展词表中每一个触发词用改进的 TF*IDF 的方法算出 *Score*
- 2、根据 *Score* 设定一个阈值,阈值一般比较高,使得大于 *Score* 的基本上都是单义词
- 3、对于含有大于阈值的那些词(少部分)的候选事件,直接查表确定事件的类型与子类型。
- 4、对于含有小于阈值的那些词(大部分)的候选事件,分别使用 ME/SVM 两种机器学习的方法来判断该候选事件是否是所规定的事件。

这样,对于大于阈值的触发词,一般为单义词,出现次数不多,但是只要出现所在候选事件就是所规定的事件,如果含有这类词的候选事件使用机器学习的方法来识别,由于实例较少,所以不具典型性,易识别错误,导致召回率降低。而对于小于阈值的触发词,一般都有一词多义现象,因此使用直接查表的方法很容易多标出很多事件,导致精确率降低,因此使用机器学习的方法来解决。

4 实验结果及分析

4.1 实验数据及评测标准

我们使用 ACE2005 年的训练数据作为实验数据。数据的来源为 Broadcast News, Newswire 和 Newspaper,共 633 篇,随机选取其中 473 篇(3/4)作为训练语料,160 篇(1/4)作为测试语料。ACE 评测的训练数据,不但标注了实体以及实体的各种属性,还标注了事件及事件的一些属性如:触发词,类别,子类别等等。从训练语料中提取出来的触发词共有 653 个,当扩展阈为 $n=2$ 时,可用《同义词词林(扩展版)》扩展出 1426 个词,并生成含有触发词的实例(句子)10014 个,其中正例 2378 个,反例 7636 个;当扩展阈为 $n=3$ 时,扩展出 946 个词。

对于事件类别识别的性能评价,采用 F 值(F-Score)对最终系统的性能进行评价。定义如下:

$$F - score = \frac{2PR}{P + R} \quad (3)$$

其中 P 为准确率, R 为召回率, 其定义分别为:

$$P = \frac{\text{找到的正确的事件个数}}{\text{找到事件总数}} \quad (4)$$

$$R = \frac{\text{找到的正确的事件个数}}{\text{正确的事件总数}} \quad (5)$$

4.2 实验结果

我们分别用基于触发词词表的方法和机器学习的方法进行了对比实验, 其中, 仅基于词表的方法删去 50 个 $Score$ 较低的词, 并使用扩展阈为 2 和 3 的同义词词林对触发词进行扩展; 机器学习的方法使用了最大熵(ME)和支持向量机(SVM)两种分类器, 并基于扩展阈为 2 的扩展词表。其中最大熵的迭代次数 i 取 60; SVM 实验中选取了 SVMlight 工具包, 取触发词的权重为 2, 其他特征的权重为 1, 取 $d=2$, 即二次多项式 Kernel, 并取 $j=1.35$ 。实验结果见表 1。

表 1 基于触发词表和使用机器学习两种方法的实验结果对比

Tab.1 Comparison of the experiment results of the approach based on trigger words only and using machine learning

实验方法	P	R	F
词表 Score 筛选	46.6%	70.6%	56.2%
扩展词表($n=2$)	24.3%	84.4%	37.8%
扩展词表($n=3$)	39.2%	72.5%	50.9%
ME	70.7%	65.2%	67.8%
SVM	74.3%	65.6%	69.7%

从上面的实验结果可以看出, 使用机器学习的方法可以有效地在由触发词表直接抽取出的候选事件中, 区分出真正的事件。它虽然导致了召回率的少许降低, 但大幅度的提高了精确率, 提高了整个抽取系统的性能。而基于扩展词表的方法召回率非常高, 弥补了机器学习的召回率低的问题。

在基于扩展的触发词表和机器学习结合的方法中, 根据扩展后触发词词表的分数的分布规律, 取阈值为 0.9, 由扩展的触发词词表和机器学习的两种方法 ME、SVM 相结合, 得到的结果见表 2。表 2 扩展触发词表和机器学习相结合的实验结果

Tab.2 Experiment result of machine learning combining with selection by score of trigger words

实验方法	P	R	F
触发词表+ME	70.3%	68.1%	69.2%
触发词表+SVM	72.6%	67.0%	69.9%

由实验结果可以看出, 基于扩展的触发词表的方法在解决 $Score$ 值较大, 即单义词的时候很有效, 因为这只依靠触发词就能识别出是否是某类事件; 但是当 $Score$ 值较小时, 即存在一词多义现象时, 就不能单单依靠触发词来识别, 而要加入其它的特征进行学习来判定, 即使用机器学习的方法。因此这两种方法扬长避短, 达到较为理想的效果。

由于 ME 是自动学习各特征的参数, 而 SVM 可以手工设定特征的参数, 并且适合于小规模语料, 因此 SVM 的 F 值比 ME 高出了 0.7%。而且我们能够看出, ME 和 SVM 这两种机器学习的方法在特征一定的情况下, 最终的效果相差不多, 说明对于不同的分类器, 特征选取才是最重要的。

5 结论和展望

本文结合 ACE 评测介绍了信息抽取领域中的事件抽取工作的主要内容, 重点研究了事件类别的确定方法。通过大量实验, 本文提出了用同义词词林扩展触发词词表和机器学习的方法 (ME/SVM) 相结合来确定事件的类别, 该种方法的优点是有效的利用了汉语中触发词和上下文对事件存在性和类型的决定作用, 与传统的使用模版的方法相比, 能够更快的扩展到其它领域。实验表明, 基于触发词词表和两种机器学习 (ME/SVM) 相结合的方法使得最终的 F-Score 分别达到了 69.2%和 69.9%, 达到了较为理想的效果。当然, 目前还有一些问题尚未解决, 比如: 同一触发词对应多个事件类别的问题, 如何寻找更好的特征, 改善分类器的性能也是我们今后工作的目标。

参考文献:

- [1] ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Events, National Institute of Standards and Technology, 2005
- [2] Rohini Srihari, Wei Li. Information Extraction Supported Question Answering. Proceedings of TREC-8. 15 October, 1999.
- [3] Michael White, Tanya Korelsky, Claire Cardie. Multi-document Summarization via Information Extraction. First International Conference on Human Language Technology Research (HLT), 2001.
- [4] 车万翔, 刘挺, 李生. 实体关系自动抽取. 第一届全国内容安全与信息检索学术会议, 2004.
- [5] Milena Yankova. Focusing on Scenario Recognition in Information Extraction.
- [6] Chang-Shing Lee, Yea-Juan Chen, and Zhi-Wei Jian. ONTOLOGY-BASED FUZZY EVENT EXTRACTION AGENT FOR REAL-TIME CHINESE E-NEWS SUMMARIZATION
- [7] Ralph Grishman. Research in Information Extraction: 1996-98. TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland. X98-1012. October 13-15, 1998
- [8] Adam L. Berger, Vincent J. Della Pietra, Stephen A. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. Computational. Linguistics, 1996, 22(1): 39-71.
- [9] Hai Leong Chieu, Hwee Tou Ng. A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text. Proceedings of the 18th National Conference on Artificial Intelligence, 2002, pages 786--791.