半监督中文事件抽取中的模板过滤和转换方法

徐 霞[1] 李培峰[2] 朱巧明[2]
(苏州大学计算机科学与技术学院 苏州 215006)[1]
(江苏省计算机信息处理技术重点实验室 苏州 215006)[2]

摘 要 事件模板是指导事件抽取工作的依据,半监督方法下模板的准确性显得尤为重要。目前,基于双视图的"触发词-论元"模板的中文信息事件抽取系统不能有效地解决触发词一词多义的现象和模板稀疏现象。提出了一种借助论元进行触发词语义消歧的方法,并利用该方法进行模板过滤以消除无效模板的影响。另外,针对几种特殊的中文句型,根据句法结构提出了模板转换规则,从而提高了模板的适用性。在 ACE2005 中文语料上的测试表明,该方法可有效地提高半监督中文信息事件抽取系统的性能。

关键词 事件抽取,模板过滤,模板转换
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.052

Pattern Filtering and Conversion Methods for Semi-supervised Chinese Event Extraction

XU Xia[1] LI Pei-feng[2] ZHU Qiao-ming[2]
(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)[1]
(Province Key Laboratory of Computer Information Processing Technology of Jiangsu, Suzhou 215006, China)[2]

Abstract The accuracy of event patterns is very important in semi-supervised event extraction. Currently, semi-supervised Chinese event extraction system based on the pairwise pattern (e.g., Trigger-Argument) suffers much from the issues of polysemy of triggers and sparse patterns. This paper put forward a argument-based mechanism to solve trigger sense disambiguation, and then applied it to pattern filtering to eliminate invalid patterns. In addition, for several special Chinese sentence structures, this paper proposed a pattern conversion method based on syntactic structure to enhance the applicability of the pattern. The experimental results on the ACE 2005 Chinese data show that our methods can effectively improve performance of semi-supervised Chinese event extraction system.

Keywords Event extraction, Pattern filtering, Pattern conversion

在自然语言中,触发词(Trigger)跟与之关联的论元(Argument)可构成基本事件模板。尽管触发词是表征事件的最好特征,但实际上有些触发词存在着一词多义的现象。有时,单从词本身无法确定其是否是某个类型事件的触发词,那么可参照与其有关系的上下文,特别是上下文中的其他实词。自然语言的语义由名词和动词等实词的词义及结构关系决定,实词之间起着相互限制词义的作用。例如:"打"既可触发"Attack"类事件,也可触发"Phone-Write"类事件。只有明确"打"的受事者,才可明确它所触发的事件,"打人"代表"Conflict/Attack"类事件,而"打电话"代表"Contact/Phone-Write"类事件。依据此思想,为了提高种子模板的准确率,我们引入基于触发词的论元过滤机制,借助论元对种子模板进行过滤。

另外,格语法的基本思想是任何一个句子都有一个深层结构,而这个结构是以某个动词与一个或多个名词词组组成,每个名词词组与动词都以一定的句式相连。在汉语句子的实际应用中,这些名词词组与动词可能出现多种句式。这种汉语句式的多样性增加了句法成分分析的难度。例如,"歹徒打了张三"、"歹徒把张三打了"、"张三被歹徒打了"、"被歹徒打了的张三"这 4 句话,多种句式使得名词词组与动词出现的顺序不一致,导致它们的句法成分也不一样。但是从语义上来看,这些名词词组充当相同的句法成分,"歹徒"都是"施事者","张三"都是"受事者","打"是"动作"。句式的多样性会引入大量的稀疏模板(出现频度低的模板),给事件抽取带来干扰。本文从语言学角度出发,提出事件模板转换规则,对汉语的特殊句式进行转换,统一事件模板,加强事件模板的代表性。

本文第 1 节介绍半监督事件抽取的相关工作;第 2 节描述事件模板过滤方法;第 3 节分析汉语句式的多样性并给出事件模板的转换规则;第 4 节汇报实验结果与分析;最后给出了结论。

1 相关工作

目前,大多数半监督事件抽取的研究集中在英文,主要有基于分类器和事件模板两种方式。基于分类器方法通常需要

到稿日期:2014-03-25 返修日期:2014-06-17 本文受国家自然科学基金(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(11KJJ520003)资助。

徐 霞(1989—),女,硕士生,主要研究方向为中文信息处理,E-mail:xuxia1125@163.com;李培峰(1971—),男,副教授,主要研究方向为中文信息处理;朱巧明(1963—),男,教授,主要研究方向为中文信息处理。

人工标注一部分语料，而模板的方法则需要高质量的种子模板，需要为每个类型的事件定义少量的种子模板。

Riloff[1]首先将未标注的文档分成相关文档和不相关文档，然后根据已标注的文档和一系列启发式规则得到更多的事件模板识别事件。Yangarber 等用自举(Bootstrapping[2])方法实现了一个以文档为中心的半监督事件抽取系统，文档相关度的方法假定相关文档总是包含一些相关模板，它们可以用来抽取事件或确定事件类型。其中，事件模板是通过最初的种子模板以自举的方式来扩充模板。Yangarber 引入多个学习器[3]到自举程序中，进一步完善文档相关度的方法，并用多个学习器的组合来最终决定事件类型。Huang 和 Riloff[4]采用语义角色的名词[5]作为种子触发词，从相关文档中抽取事件模板来标记实例，然后得到事件抽取系统的训练器。

Stevenson 和 Greenwood[6]提出了一种模板相似度的方法，通过已有的种子模板来筛选候选模板。通常情况下，文档相关度的方法很容易选中相关文档中高频率的伪模板。为了解决这个问题，Liao 和 Grishman 将模板相似度的衡量值作为一个过滤器[7]引进到文档相关度的方法中，以消除一些伪模板。Liao 和 Grishman 进一步以信息检索机制[8]来发现相关文档并提出自学习策略。

半监督中文事件信息抽取的研究比较晚，Chen 和 Ji 借助 100 篇文档中的 500 多个标注事件，使用跨语言的多种特征方法[9]在自举过程抽取中文事件，事件抽取的 F1 值只达到 35%，这表明半监督下的中文事件抽取还存在巨大挑战。

除了上述研究，还有一些事件模板表达形式方面的研究，例如 Chambers 和 Jurafsky 的主谓、动宾结构[10,11]，Yangarber 和 Balasubramanian 的主谓宾结构 SVO[2,12]，Sudo 的模板链 chain[13]、子树结构 subtree[14]以及 Liu 和 Strzalkowski 的复杂模板结构[15]。

2 事件模板过滤方法

自然语言中句子的语义通常是由主语、宾语和谓语等实词的词义及结构关系决定，实词之间起着相互限制词义的作用。我们的基准实验仅从触发词本身来决定语义，没有对“一词多义”的现象实现词义消歧，从而导致抽取了大量假事件实例。例如，“我打电话”、“我打比赛”、“我打人”这些语句都是以“打”为动作，词义却不一样，所属事件类别也不一样。如果忽略触发词的多义性，随着自举算法迭代次数的增加，越来越多的负例会被加入。

虽然仅分析触发词本身不能识别词义，但是可以通过“打”的受事者(宾语成分)来判断是否为某类事件。上述例子中，“打”的主语都是“我”，没有明显区分度，但是很容易看出，如果受事者是 PER 类实体，“打”则是“Attack”类事件，即“我打人”是“Attack”类事件。本文将从动宾结构这个角度对动词触发词进行词义消歧。

本文以 Injure、Die、Attack 3 类事件为研究对象。根据这 3 类事件的定义可发现一个语言现象：目标事件中动宾结构的宾语一般都是人(PER)、政治性实体(GPE)或机构(ORG)。为提高种子模板的准确率，可以通过判断动宾结构中宾语的实体类型来过滤无效事件模板，避免无效事件模板的扩充，解决触发词一词多义的现象。

事件触发词、论元及其相关特征构成了事件模板的框架，本文的事件模板组成包括：触发词、论元、词性、(子)类别、依存路径，为模板定义一个六元组 Pattern=(Trigger, Trigger-POS, Arguement, Enttype, Subenttype, Path)。事件模板过滤规则方法是判断 Path 中是否含有 dobj 的事件模板的论元的 Enttype，若没有 PER、GPE、ORG 类型的论元，则其不被选为种子模板。

3 事件模板转换方法

事件过滤方法要求能够准确找出句子中的动宾结构，但是汉语句式的多样性使得现有的句法分析器不能准确给出语句的各个句法成分。此外，句式的多元化会产生大量稀疏的候选事件模板，稀疏事件模板相比于高频率事件模板很难被识别，不利于事件的识别。如果可把这些表述不同的句子转换为统一的模式，那么统一后的事件模板将更具有一般性与代表性。于是，本文提出事件模板转换方法，对事件模板进行归纳，使它们更具有代表性，以弥补汉语句式多样性带来的不足。

无论汉语句式结构多么自由，其中蕴含的最基本规则都相对稳定。本文针对语料库中大量出现的“的”、“把”、“被”型语句，从汉语语句最本质的特征出发，宏观上总结出事件模板转换方法。含有“VV-的”、“把”、“被”型语句的依存结构中，触发词与论元之间的依存关系会存在偏差。本文分析已有的依存关系，结合汉语句式本身重新确定基本事件中动词的主语、宾语、间接宾语等句法成分。该方法具有一定的通用性，与领域无关。

3.1 “VV-的”型转换方法

现代汉语中，动词后面紧跟“的”字的形式主要有两种。例如，例(1)、例(2)中的动词“受伤”、“逮捕”都修饰名词充当定语；例(3)、例(4)中动词“去世”、“杀害”与“的”连用，独立充当名词性成分。

例(1)：受伤的乘客

例(2)：被逮捕的歹徒

例(3)：特鲁多是因病去世的

例(4)：阿卡维是被犹太闹事者杀害的

从上述例子可以看出，“VV-的”修饰名词时充当定语，否则通常与“是”搭配。此外，“VV-的”常常会与“被”一起出现(本文将形如“被 * * * VV-的”的结构归为“VV-的”型)，是否存在被动关系会直接影响名词是动作的施事者还是受事者。“VV-的”型转换时如果动词存在被动关系，那么名词是动作的受事者；否则是动作的施事者。上述两种“VV-的”型转换结果如图 1 所示。

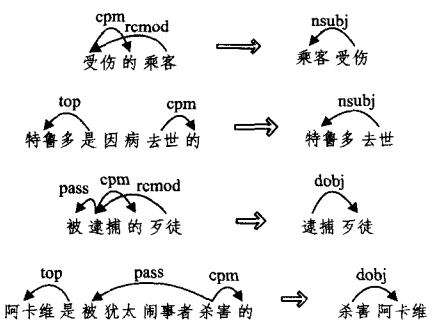

图 1 “VV-的”型转换方法

3.2 “把”型转换方法

现代汉语中，“把”字语句的形式特点是宾语提到动词与“把”之间，依据动词后面是否有补充分成两类。例(5)中动词“扔”后面没有补充成分；例(6)中动词“押”后面有动作的补充成分，这里是动作的目的地点。

例(5)：将发臭的鸭蛋朝门口扔

例(6)：把船长押到后面的机舱

从上述例子可以看出，“将/把”型句式中动词的宾语都被提到动词与“把”之间，而现有的依存分析会认为宾语是动词的施事者。“把”型转换方法是将宾语置后，成为动词的受事者。上述两种“把”型转换结果如图 2 所示。

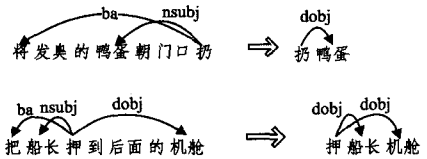

图 2 “把”型转换方法

3.3 “被”型转换方法

现代汉语中，“被”字语句的形式特点是宾语提到“被”前面，依据动词后面是否有补充分成两类。例(7)中动词“逮捕”后面没有补充成分；例(8)中动词“称”后面有动作的补充说明成分。

例(7)：歹徒被警察逮捕了

例(8)：他被南斯拉夫媒体称作总统当选人

从上述例子可以看出，“被”型句式中动词的宾语都被提到“被”前面，而现有的依存分析会认为宾语是动词的施事者。“被”型转换方法是将宾语置后，成为动词的受事者。上述两种“被”型转换结果如图 3 所示。

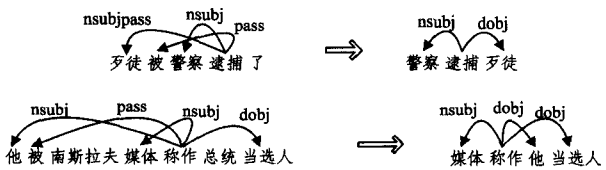

图 3 “被”型转换方法

4 实验结果与分析

本文的实验数据是 ACE 2005 中文语料，该语料将事件的类别主要分为 8 个大类及 33 个子类。对语料库中 633 个中文文本统计表明，Injure 事件有 163 个，Die 事件有 243 个，Attack 事件有 534 个。本实验主要通过高准确率的种子模板来识别具有代表性的 Injure、Die、Attack 3 类事件，以汇报准确率 P、召回率 R 和 F1 值作为评价标准。

4.1 基准系统

基准实验主要用自举方法实现半监督的中文事件抽取，首先以少量种子模板为基础，分别采用文档相关度方法与语义相似度方法来扩充种子模板，为事件抽取服务，发现它们分别带来高准确率与高召回率的不同优势；然后提出双视图互训练方法，即在文档相关度与语义相似度两个视图下择优一同影响种子模板，吸收双方的优点，共同发现目标事件。从表 1 可知，基准系统采用的双视图方法使 F1 值达到了 52.1%，取得了不错的效果。

实验中的中文文本以句子为单位，利用苏州大学自然语言处理平台集成的工具进行分词、词性标注、实体类别识别。在此基础上进行句法和依存分析，得到依存和句法分析树。将词性为动词或名词的词汇作为触发词，依次与该句中的实体形成触发词-论元式，并记录论元的类别、依存关系等特征，得到本文的候选模板。

系统最初通过种子触发词得到种子模板，将候选模板中包含种子触发词的模板选为种子模板。实验中选用能代表特定类型事件的最少量的触发词，具体如下：

- Injure 事件：“伤”
- Die 事件：“死”
- Attack 事件：“攻击”、“冲突”、“打”

但是基准系统没有解决一词多义的现象，将“打电话”、“打篮球”、“打比赛”与“打人”一同认为是“Attack”事件，导致引入大量的假事件实例。同时，多元化的汉语句式会得到大量的稀疏模板，给事件抽取带来困难。

4.2 实验结果与分析

针对触发词一词多义现象和稀疏模板现象，本文在基准系统上加入事件模板过滤和转换方法。表 1 给出了事件模板过滤和转换方法对系统性能的影响，与基准实验相比，准确率 P、召回率 R 和 F1 值均有不同程度的提高，取得了很好的进步。

表 1 实验结果

实验	P(%)	R(%)	F1(%)
双视图的方法	69.8	41.5	52.1
+转换方法	73.5	43.0	54.3
+转换方法+过滤方法	77.6	43.1	55.4

语料库里大量“VV-的”、“把”、“被”句式不仅会影响句法分析，而且会带来大量的稀疏事件模板，相比于高频率事件模板，稀疏事件模板很难被识别。基于语言学理论基础，结合汉语语句的特征，对候选模板中 948 个“VV-的”型、168 个“把”型、254 个“被”型句式的事件模板进行转换，转换方法的准确率(能正确找到动词的受事者的模板个数/该句式的所有模板)如表 2 所列，数据表明本文的转换方法具有一定的通用性，与领域无关。

表 2 转换方法的准确率

句式	模板数量	正确转换数量	准确率(%)
VV-的	948	796	84.0
把	168	136	81.0
被	254	219	86.2

从表 1 可知，转换方法使得抽取的正例 TP 的数量增加了 3.6%，负例 FP 的数量减少了 13.8%。双视图方法在筛选候选模板过程中的转换方法增加了事件模板的统一性，可以有效地让更多的正例模板的排名次序优于负例模板，从而识别出更多的目标事件。表 3 给出了模板转换方法对特殊句式与其它句式事件抽取结果 P、R 和 F1 值的变化值。总体来说，模板转换方法可以提高事件抽取的性能。但本文的事件抽取依赖种子模板，被抽取的含特殊句式的事件存在瓶颈。对特殊句式而言，模板转换方法可以提高事件抽取的准确率，但未能提高召回率。

表 3 模板转换方法对事件抽取的影响

句式	ΔP(%)	ΔR(%)	ΔF1(%)
特殊句式	+17.5	−1.2	+3.8
其它句式	+2.6	+1.7	+2.1

(下转第 262 页)

所以，系统 $x(t)$ 是全局指数稳定的，稳定时间估算为

$$T=t_0+\frac{V^{1-\frac{1+\eta}{2}}(t_0)}{2\lambda(1-\frac{1+\eta}{2})}$$

基于定理 1，同样可以推导出以下推论。

推论 1 如果存在正常量 s_1 满足

$$k\geqslant(\lambda_{\max}(A+A^{\mathrm{T}})+s_1\lambda_{\max}(B+B^{\mathrm{T}})+s_1^{-1}L_{\max}^2)/2 \quad (10)$$

那么，系统 $x(t)$ 是全局指数稳定的，同步时间估算为

$$T=t_0+\frac{V^{1-\frac{1+\eta}{2}}(t_0)}{2\lambda(1-\frac{1+\eta}{2})}$$

注：从推论 1 可知，对于更小的 λ，稳定时间 T 将变得更大，相应地，对于合适的 $V(t_0)$ 和 η，稳定时间 T 随着 λ 增长而减小。

结束语 采用本文提出的控制方法，不仅能够很好地抑制忆阻混沌系统混沌现象的产生，而且使得受控系统实现了有限时间稳定控制。理论分析显示了该方法的正确性和有效性。

参 考 文 献

[1] Itoh M, Chua L O. Memristor oscillators[J]. International Journal of Bifurcation and Chaos, 2008, 18(11): 3183-3206

[2] Muthuswamy B, Kokate P P. Memristor Based Chaostic Circuits [J]. IETE Technical Review, 2009, 26(6): 415-426

[3] 包伯成，王其红，许建平. 基于忆阻元件的五阶混沌电路研究[J]. 电路与系统学报，2011，2(16)：66-70

[4] 张向华. 一种改进的基于时空混沌系统的 Hash 函数构造方法[J]. 计算机科学，2009，36(7)：252-255

[5] 柴秀丽，王玉璟，袁光耀，等. 未知干扰下混沌系统的修正函数投影滞后同步[J]. 计算机科学，2014，41(4)：283-286，301

[6] 吴凌燕，孙永芹. MATLAB 在电路求解中的运用 [J]. 自动化与仪器仪表，2013(05)：122-123

[7] 王清华，宋卫平，宇文雄. 忆阻器特性及其在电路设计中的应用[J]. 电子元件与材料，2014(3)：5-8

[8] 李伟，熊静，李春成. 一个新混沌系统的分析及电路实现 [J]. 内江师范学院学报，2014(2)：25-30

[9] 徐伟，马进颖. 蔡氏混沌电路在 Multisim 软件中的设计与仿真[J]. 电子器件，2013(6)：904-909

[10] 孙克辉，贺少波，朱从旭，等. 基于 C0 算法的混沌系统复杂度特性分析 [J]. 电子学报，2013(9)：1765-1771

[11] 杨留猛，俞建宁，安新磊，等. 一个新三维自治系统的混沌分析及电路模拟[J]. 重庆理工大学学报，2012，26(12)：127-133

(上接第 255 页)

针对一词多义的现象，加入事件模板过滤机制，根据“动宾结构”中论元的实体类型来判断触发词的词义，过滤了 460 个高概率易被抽取的事件，其中仅 8 个正例。实验表明，过滤方法可以很好地过滤“攻击的后果”、“打理日常工作”等常被抽取的负例，使得抽取的错误正例 FP 的数量降低 19.4%，准确率达到 77.6%，F1 值达到 55.4%，进一步改善了系统的性能。

结合事件模板过滤和转换两个方法，可以很好地提高中文事件抽取系统的性能，但是不能摆脱对种子模板的依赖，因此还存在很大的发展空间。

结束语 本文主要针对触发词一词多义现象和模板稀疏现象对中文事件抽取系统性能的影响，提出事件模板过滤和转换方法，借助触发词的论元进行语义识别，过滤无效模板，减少假事件实例；且针对中文的特殊句型，根据句法结构提出了模板转换的规则，从而提高了模板的适用性，分别提高了种子模板的准确性与事件模板的统一性，很好地优化了中文事件抽取系统的性能。

在接下来的工作中，我们将摆脱对种子模板的依赖，从事件推理的角度识别事件，进一步改善中文事件抽取工作。

参 考 文 献

[1] Riloff E. Automatically Generating Extraction Patterns from Untagged Text[C]// Proceedings of the Thirteenth National Conference on Artificial Intelligence. 1996: 1044-1049

[2] Yangarber R, Grishman R, Tapanainen P, et al. Automatic Acquisition of Domain Knowledge for Information Extraction[C]// Proceedings of the 18th Conference on Computational linguistics. 2000: 940-946

[3] Yangarber R. Counter-Training in Discovery of Semantic Patterns[C]// Proceedings of ACL 2003. 2003: 343-350

[4] Huang Rui-hong, Riloff E. Bootstrapped Training of Event Extraction Classifiers[C]// Proceedings of EACL 2012. 2012: 286-295

[5] Phillips W, Riloff E. Exploiting Role-Identifying Nouns and Expressions for Information Extraction[C]// Proceedings of RANLP 2007. 2007: 468-473

[6] Stevenson M, Greenwood M. A Semantic Approach to IE Pattern Induction[C]// Proceedings of ACL 2005. 2005: 379-386

[7] Liao Sha-sha, Grishman R. Filtered Ranking for Bootstrapping in Event Extraction[C]// Proceedings of COLING 2010. 2010: 680-688

[8] Liao Sha-sha, Grishman R. Can Document Selection Help Semi-supervised Learning? A Case Study on Event Extraction[C]// Proceedings of ACL 2011. 2011: 260-265

[9] Chen Zheng, Ji Heng. Can One Language Bootstrap the Other: A Case Study on Event Extraction [C] // Proceedings of HLT-NAACL 2009 Workshop on Semi-supervised Learning for Natural Language Processing. 2009: 66-74

[10] Chambers N, Jurafsky D. Unsupervised Learning of Narrative Event Chains[C]// Proceedings of ACL-HLT 2008. 2008: 787-797

[11] Chambers N, Jurafsky D. Unsupervised Learning of Narrative Schemas and Their Participants[C]// Proceedings of ACL 2009. 2009: 602-610

[12] Balasubramanian N, Soderland S, Mausam, et al. Generating Coherent Event Schemas at Scale[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1721-1731

[13] Kiyoshi S, Satoshi S, Ralph G. Automatic Pattern Acquisition for Japanese Information Extraction[C]// Proceedings of HLT 2001. 2001: 1-7

[14] Kiyoshi S, Satoshi S, Ralph G. An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition[C]// Proceedings of ACL 2003. 2003: 224-231

[15] Liu Ting, Strzalkowski T. Bootstrapping Events and Relations from Text[C]// Proceedings of EACL 2012. 2012: 296-305