

基于框架语义分析的社交网络药品不良事件抽取^{*}

由丽萍 王世钰 李朝翻

(山西大学经济与管理学院 太原 030006)

〔摘要〕 目的/意义 基于社交网络评论文本抽取药品不良事件,为药品研发和安全监管提供参考。方法/过程 采用框架语义理论,结合《监管活动医学词典》术语集构建药品不良事件分类词表;基于词典和规则匹配的方法识别事件类别和框架元素,利用语义信息实现药品不良事件框架填充。结果/结论 选取社交网络药品评价实例进行药品不良事件信息抽取可行有效,有助于框架语义分析方法在医疗专业领域的深度应用和价值实现。

〔关键词〕 社交网络;框架语义;药品不良事件;信息抽取;自然语言处理

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2023.07.009

Extraction of Adverse Drug Events from Social Media Based on FrameNet Semantic Analysis

YOU Liping, WANG Shiyu, LI Chaofan

College of Economics and Management, Shanxi University, Taiyuan 030006, China

〔Abstract〕 **Purpose/Significance** To extract adverse drug events based on the text of social network comments, and to provide references for drug research and safety regulation. **Method/Process** The FrameNet semantic theory is adopted and combined with the *Medical Dictionary of Regulatory Activities* terminology set to construct the adverse drug event classification lexicon. The lexicon and rule matching-based approach is used to identify event categories and frame elements, and semantic information is used to achieve the filling of adverse drug event frame. **Result/Conclusion** The selection of social network drug evaluation examples for adverse drug event information extraction is feasible and effective, which contributes to the in-depth application and value realization of the FrameNet semantic analysis method in the medical field.

〔Keywords〕 social network; frame semantics; adverse drug event; information extraction; natural language processing (NLP)

1 引言

近年来,药品安全性事件频发,对公众健康造

成严重威胁,药品安全问题逐渐引起舆论热议。据世界卫生组织统计,全球死亡患者中有 1/3 死于用药不合理^[1]。在我国,药疗事故占医疗事故的 30%,药品不良事件已成为严重的公共卫生问题。随着信息技术在医疗领域的广泛应用,在线医疗平台和药品电商等互联网产业蓬勃兴起,积累了大量药品咨询记录和药效评价信息。抽取此类信息不仅对分析患者用药后的反馈具有重要意义,而且可以为药品生产、经营、使用及药政部门管理提供参

〔修回日期〕 2022-10-29

〔作者简介〕 由丽萍,教授,硕士生导师,发表论文 30 余篇。

〔基金项目〕 国家自然科学基金面上项目(项目编号:61772324)。

考，以提高监管力度，推动社会卫生事业进步。

目前，事件抽取相关研究成为国内外自然语言处理领域的重要研究方向之一，旨在从非结构化文本中抽取公众感兴趣的事件信息，并以结构化形式呈现出来。已有研究立足于从信息^[2]、事件^[3]、关系^[4]3 个层次展开，将相关事件提取技术应用于特定领域，如政治外交^[5]、法律^[6]以及临床^[7]等，以提升特定领域事件抽取精确率。在生物医学领域，陈瑶等^[8]基于深度学习模型，引入注意力机制，对字向量与分词向量进行优化，研究异常检出率报告中的“药品-不良反应”关系抽取。张玉坤等^[9]提出一种联合神经网络模型，将实体识别和关系抽取联合起来，研究中文医疗关系抽取。Kilicoglu H^[10]和 Bui Q C^[11]分别提出 ConcordU 和 BioSEM 生物医学事件抽取方法。这些基于词向量的方法可操作性强，但缺乏词汇资源的有效组织和语义关联描述，表示深度及语义理解不足，处理精度不高，应用价值受限；而基于语义分析则能满足对自然语言深层语义理解的需要，准确识别不良反应与感知部位之间的语义关系。

本研究基于框架语义理论，结合《监管活动医学词典》(Medical Dictionary for Regulatory Activities, MedDRA)^[12]术语集构建药品不良事件分类词表；根据词典和句法规则分析公众在社交网络中对药品的评论文本，识别药品不良事件相关语义信息，填充药品不良事件框架；基于框架语义分析结果，完成药品不良事件抽取。这是情感分析向医疗健康领域纵深发展的一次有益探索，兼具社会意义与学术价值。

2 研究方法

药品不良事件抽取由文本预处理、框架语义标注和药品不良事件框架填充 3 部分构成。首先，以微博、知乎等社交网络平台为数据来源，通过网络爬虫技术，抓取包含有“副作用”“药品不良反应”等信息的用药反馈评论文本，并对其进行分词、词性标注等预处理。其次，基于框架语义标注规则对文本进行框架元素标注，并通过匹配药品不良事件

分类词表及程度值计算规则，将语义识别出的谓词匹配不良事件触发词，确定药品不良事件及其整体程度值。最后，根据框架语义分析结果填充药品不良事件框架，从而完成药品不良事件抽取，见图 1。

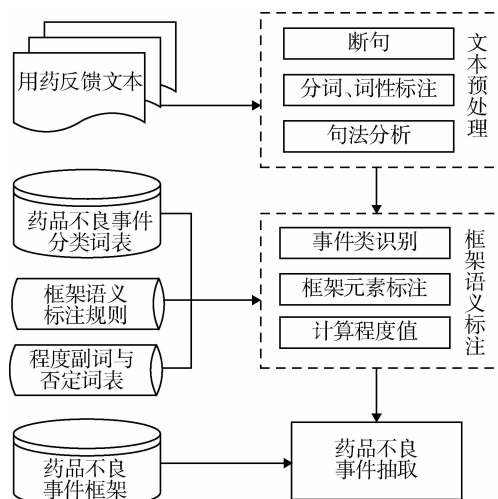


图 1 药品不良事件抽取流程

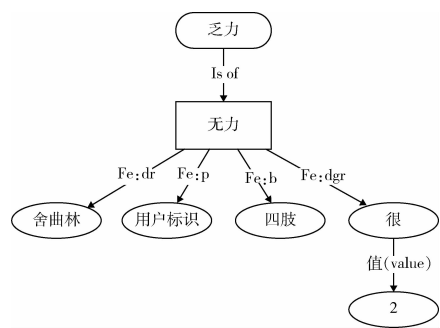
3 框架语义分析

3.1 框架语义表示模型

框架语义理论是美国语言学家 Fillmore C J^[13]提出的一种以框架为单位描述词语意义以及句子语义结构的方法体系。框架是语言使用者在其认知经验中存储的关于词语和句子语义理解的图示化情景，由一系列框架元素组成，表明情景中的动作主体、客体、程度等其他概念角色。以框架语义模型表示药品不良事件，即将语言表达中的谓词和语义角色标注为相应的事件类别和患者、药品、感知部位等框架元素，其定义如下：

$$E_i (dr_i, p_i, b_i, dgr_i) \quad (1)$$

其中， E_i 为句子中一个短语 i 的事件类； dr_i 为药品名称； p_i 为患者； b_i 为感知部位； dgr_i 为程度值。程度值表示药品不良事件的严重程度，依据医学程度常用分级，进行 4 级区分，1~4 级依次为轻度、中度、重度、极重度。如评论文本“吃了舍曲林感觉四肢很无力”，其模型为： E_i : 乏力 (dr_i : 舍曲林, p_i : 匿名用户, b_i : 四肢, dgr_i : 2)，见图 2。



3.2 药品不良事件分类词表构建

在药品不良事件框架中，事件类是由句子中

的谓词（动词和形容词）所激活的语义内容，其在框架中位于核心支配地位，与框架元素相关联。本文以 MedDRA 为基础确立基本事件类别，在爬取的大量社交网络文本中提取谓词与 MedDRA 进行匹配，构建药品不良事件分类词表。MedDRA 具有层级结构，从低到高分别为低位语（lowest level term, LLT）、首选语（preferred term, TP）、高位语（high level term, HLT）、高位组语（high level group term, HLT-GT）和系统器官分类（system organ class, SOC），其中术语即为低位语，同义术语或等同术语归于首选语，见表 1。

表 1 药品不良事件分类词表示例

高位语	首选语	低位语	评论词
意识障碍（disturbances in consciousness）	嗜睡（somnolence）	昏昏沉沉	昏昏沉沉、昏沉、头昏、迷迷糊糊、迷糊
各种睡眠（dyssomnias）	嗜睡（somnolence）	昏昏沉沉	昏昏沉沉、昏沉、头昏、迷迷糊糊、迷糊
各种颤抖（tremor（excl congenital））	震颤（tremor）	颤抖	抖、颤抖、颤栗、打颤、战栗
入睡和睡眠障碍 （disturbances in initiating and maintaining sleep）	失眠（insomnia）	失眠	失眠、睡不着、翻来覆去、辗转反侧、夜不能寐
各种表皮和真皮病症 （dermal and epidermal conditions）	感觉减退 （hypoesthesia）	麻木	麻木、麻麻的、酥麻、发麻、麻酥酥
神经学症状和体征 （neurological signs and symptoms）	感觉减退 （hypoesthesia）	麻木	麻木、麻麻的、酥麻、发麻、麻酥酥

3.3 标注规则

用药反馈文本中句法依存关系与语义角色之间有较强的对应规律，因此，基于规则匹配的方式识别句子中的感知部位，即谓词所支配的语义角色。框架语义标注预设规则为：

$$LU_{frame} [SBV \oplus ATT - head_{theme}] [ADVd_{degree}] [AD - Vn_{negative}] \dots$$

(2)

其中，LU 代表药品不良事件评论文本中的谓词，SBV 代表主语成分。由于评论句在表达方式上具有句式简单、主语省略甚至独词成句等特点，因此，除了谓词是必要成分外，主语和其他修饰语均为可选项，以“[]”表示。根据依存句法分析结果，若谓词支配一个主语成分，则该成分标注为该不良事件的框架元素（theme），即感知部位；若谓

词处于偏正结构中定语中心语（ATT - head）位置，则标注为感知部位。符号⊕表示这两种情况是“异或”的逻辑关系，即要么 SBV 成立，要么 ATT - head 成立，二者不能同时成立。若依存句法结构中出现程度副词做状语成分（ADVd），则标注为框架元素“程度”（degree）；若出现否定副词做状语成分（AD - Vn），则标注为“否定修饰”（negative），二者将作为程度值计算依据。

4 药品不良事件抽取

4.1 依存句法分析

依存句法分析是按照指定语法规则将特定序列用图结构呈现的自然语言语句，可以直接展示出句中词与词之间的句法关系。在基于图结构的表示方

法中,每个词语代表一个实体,用节点表示,词语节点之间的连线则表示其之间的句法关系。这种表现形式清晰简洁、标注容易,其句法功能标记与语义结构相互对应并有较强规律性,因此可以有效地识别句子中谓词及其所支配的语义角色,从而为实现药品不良事件评论文本的语义标注提供规则依

据。示例 1 依存句法关系,见图 3。其中 v 代表动词、a 代表形容词, n 代表名词, d 代表副词, u 代表助词, m 代表数词, p 代表介词, wp 代表标点符号。

示例 1: 刚开始服用舍曲林就恶心,没食欲,手会轻微抖动,内心感觉很麻木。

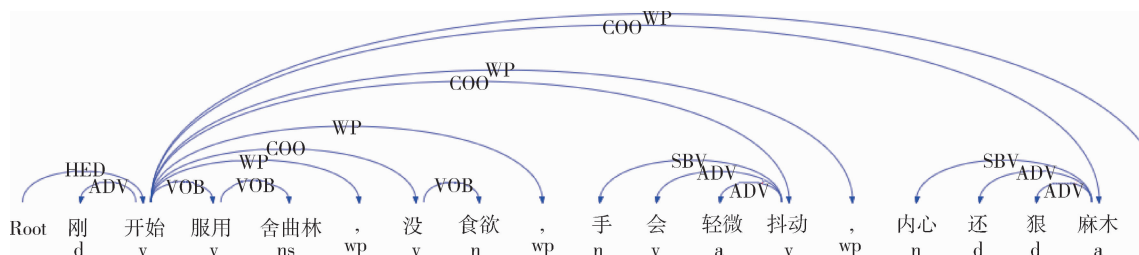


图 3 依存句法分析

4.2 药品不良事件类别识别

药品不良事件识别基于框架语义标注结果,与药品不良事件分类词表匹配事件触发词,从而识别出药品不良事件,然后进一步匹配上级类别。若只匹配到单一上级类别,直接输出药品不良事件类别;若存在多个上级类别,则存在归类歧义,通过将上下文中出现的其他词(主要为表示感知部位的词)进一步限定,从而筛选出合适的上级类别,最终完成药品不良事件类别的识别。

针对示例 1 进行药品不良事件类别的识别。首先,将动词“没食欲”、动词“抖动”和形容词“麻木”匹配药品不良事件类词表,分别对应为“食欲减退”“震颤”和“感觉减退”。其中,低位语“食欲减退”和“震颤”,分别对应高位语“各种食欲障碍”和“各种颤抖”。“感觉减退”则对应两个高位语“各种表皮和真皮病症”“感觉异常和感觉迟钝”,存在歧义。根据上下文出现的词汇“内心”可将“感觉减退”进一步限定为“感觉异常和感觉迟钝”。最终,该条评论所识别出的药品不良事件类别分别为“各种食欲障碍”“各种颤抖”和“感觉异常和感觉迟钝”。

4.3 药品不良事件框架元素识别

4.3.1 药品名称和患者识别 对网络文本数据进

行分词和词性标注,基于实体命名识别规则,从经过预处理的评论文本中识别药品名称,即对词性标注为专有名词(nz)的词语前后两个词进行扫描后匹配以下模式,若模式相符合则识别字符串中的*为药品名称。

模式 1: 吃 v 了 u * nz。例如:吃了舍曲林。

模式 2: 长期 a 使用 v * nz。例如:长期使用舍曲林。

模式 3: 服用 v * nz。例如:服用舍曲林。

模式 4: * nz 的 u 副作用 n。例如:舍曲林的副作用。

模式 5: # m 片/盒/瓶 p * nz (#表示任意数词)。例如:两片舍曲林。

针对示例 1,药品名称识别符合模式 3,于是将药品名称识别为“舍曲林”。根据各社交平台中发布相关评论的用户资源标识符填充至框架元素“患者”中。

4.3.2 感知部位识别

即谓词所支配的框架元素,根据前文框架语义分析标注规则进行识别。在示例 2 中,可将主谓结构中的主语(SBV)“嘴唇”“头”和定中结构的中心语(ATT-head)“肠胃”识别为感知部位。

示例 2: 最/d 开始 v/吃/v 舍曲林/n [感知部位 嘴唇/n] 不时/d 会/v <震颤 颤抖/v>, /wp [感知部位 头/n] 也/d 很/d <震颤 晕/a>, /wp [感知部位 肠胃/n] 的/u <疼痛和不适 不适/a> 逐渐/d 加重/v。/wp

针对前文示例 1,根据框架语义标注规则,将与“抖动”“麻木”分别构成主谓结构的中心语

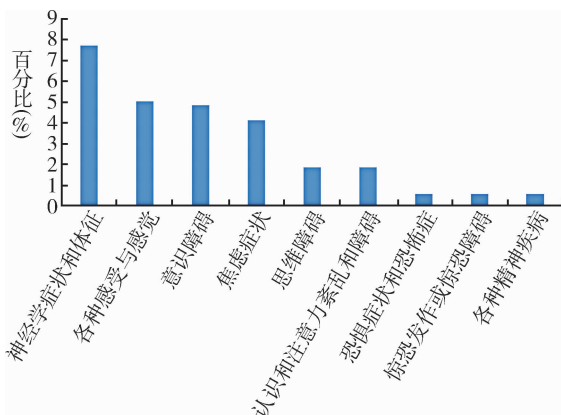


图5 影响精神反应的药品不良事件比例

由此可知,该药品最严重的不良反应为睡眠障碍,根据患者个体差异主要分别体现为失眠和嗜睡;身体上还会有虚弱状态,体现为累、乏力等,并且会恶心、呕吐;精神上则主要表现为神经学症状和体征及各种感受与感觉,主要体现在头晕、情绪差,并且会存在意识障碍、困倦迟钝等。

6 结语

本文基于社交网络文本框架语义分析研究药品不良事件抽取。基于 MedDRA 术语集构建药品不良事件框架语义分类词表,为医疗领域文本分析提供语义资源;通过框架语义理论,识别药品不良事件以及与其相关药品名称、患者、感知部位,提高事件抽取的完整性和准确性;以“盐酸舍曲林”为例,对比抽取结果与药品说明书,证实该方法的有效性。本文研究结果一方面为医疗健康领域提供语义知识库资源;另一方面为社交网络文本挖掘向医疗健康领域纵深发展提供新方向。

相比于传统问卷和量表分析方法,文本挖掘方法可以批量处理患者关于药品的评论文本,但目前所构建词表规模有限,尚未覆盖不同种类药品的不良反应,无法实现大规模不良反应识别;社交媒体数据繁杂、各平台之间信息互通,评论者可通过多平台对药品进行同一评述,可能造成数据交叉重复,存在一定局限性。在未来工作中,将尝试采用机器学习和规则相结合方式,扩充词典规模,完善识别规则,从而提高事件抽取准确率;并且尝试对

所抽取药品不良反应进行本体存储与构建,从而形成知识图谱以实现信息的深度关联与展现。

参考文献

- 1 李岚,丁文茂,李言. 武汉市某大型综合医院门诊患者参与用药安全现状调查 [J]. 中国药师, 2022, 25 (4): 664-667.
- 2 HONG F, DAN G, LI H. Extraction of spatio-temporal information of earthquake event based on semantic technology [C]. Enshi: International Symposium on Multispectral Image Processing & Pattern Recognition International Society for Optics and Photonics, 2015.
- 3 季忠祥,吴悦. 基于组合神经网络的中文事件抽取 [J]. 上海大学学报(自然科学版), 2021, 27 (3): 535-543.
- 4 彭博. 融合知识图谱与深度学习的文物信息资源实体关系抽取方法研究 [J]. 现代情报, 2021, 41 (5): 87-94.
- 5 崔莹. 基于相似义原和依存句法的政外领域事件抽取方法 [J]. 计算机工程与科学, 2020, 42 (9): 1632-1639.
- 6 LI Q, ZHANG Q, YAO J, et al. Event extraction for criminal legal text [C]. Nanjing: 2020 IEEE International Conference on Knowledge Graph (ICKG), 2020.
- 7 李井竹,陆玉婷,顾进广. 基于句法分析的临床指南事件及事件关系提取 [J]. 武汉大学学报(理学版), 2015, 61 (2): 156-162.
- 8 陈瑶,吴红,葛卫红,等. 基于深度学习模型的我国药品不良反应报告实体关系抽取研究 [J]. 中国药科大学学报, 2019, 50 (6): 753-759.
- 9 张玉坤,刘茂福,胡慧君. 基于联合神经网络模型的中文医疗实体分类与关系抽取 [J]. 计算机工程与科学, 2019, 41 (6): 1110-1118.
- 10 KILICOGLU H, BERGLER S. Syntactic dependency based heuristics for biological event extraction [C]. Boulder: North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.
- 11 BUI Q C, CAMPOS D, MULLIGEN E V, et al. A fast rule-based approach for biomedical event extraction [C]. Sofia: Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2013.
- 12 International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). MedDRA medical dictionary for regulatory activities [EB/OL]. [2022-09-18]. <http://www.ich.org/page/meddra>.
- 13 FILLMORE C J. Frame semantics and the nature of language [J]. Annals of the NY Academy of Sciences, 1976, 280 (2): 20-32.