怎样构建面向事实性表达研究的法律专题语料库?*

袁毓林[1、2]　崔玉珍[3]　孙　竞[3]　游　豪[2]

(1. 澳门大学人文学院中国语言文学系,澳门; 2. 北京大学中文系/中国语言学研究中心/计算语言学教育部重点实验室,北京 100871; 3. 中国政法大学人文学院中文系,北京 100088)

提　要　本文讨论怎样以真实的法庭审判语料为基础,构建面向事实性表达研究的专题语料库。文章指出,这种语料库的标注应该聚焦于事实性构式的三种基本信息:事实、非事实和反事实,并从构式、语段、词汇三个层次挖掘跟事实性信息相关的形式特征,最终形成了完整的标注体系和标注规范;进而采用 BMES 四元序列标注法,对这种语料进行事实性信息的标注。文章最后通过实例,展示了面向事实性信息的专题语料库的进一步应用的可能性。

关键词　事实性　专题语料库　法律语料库　反事实　论辩研究

DOI:10.16027/j.cnki.cn31-2043/h.2023.02.009

一、事实性构式

事实性(factcity)是指语句及其构成成分所表示的意义跟实际情况是否相同这样一种语义学性质(袁毓林 2020)。事实性表达和语义推理有着密切关系。袁毓林(2020)提出:"制导语义推理的许多机制就内置于语言的具体的词项或构式之中;或者说,隐含在词项和构式中的某种语义成分及其背后的概念结构中;有时甚至还有其外在的形式标记或线索,它们透露了说话人在命题的事实性或主观态度上的倾向性。"因此,一些特定的句式"就为语言推理规定了方向和设定了轨道,从而成为方便人们语言生成和理解的导航机制"。其中,事实性表达具有特定的语义预设和蕴涵,是语言推理的一种导航机制。

反事实条件句是一种典型的事实性表达,也是反事实思维最典型的语言表达形式。从语言形式上看,反事实思维通常用条件复句表达:前件是条件小句,表示一个跟实际情况相反的命题;后件是结果小句,表示在这种前提下所产生的后果(袁毓林 2015)。但是,并非所有的条

* 本课题的研究得到教育部人文社会科学研究规划基金项目"法庭立场表达与身份构建的互动研究"(项目编号:21YJA740005)和澳门大学讲座教授研究与发展基金(项目编号:CPG2022-00032-FAH)和启动研究基金(项目编号:SRG2022-00011-FAH)的资助,谨此致谢。

本文曾在 2022 年 12 月 3 日由上海外国语大学语料库研究院与复旦大学《当代修辞学》编辑部举办的"语料库运用与修辞研究发展前沿理论工作坊"上宣读。

件句都能表达反事实意义; 并且,不同类型的条件句表达反事实的能力也并不相同。比如,Wang(2012)指出,汉语违实句(即反事实句)常常发生在条件句框架内,并根据违实生成能力将假设连词分成了三类: 违实义假设连词("要不是、若非、如果不是、若不是"),可能产生违实义的假设连词("如果、要是、假如、假使、假设、倘若、如果、设若")和不能产生违实义的假设连词("万一")(转引自雍茜 2015)。李思玮(2019)也提到"如果 P,就 Q"既可以表达反事实,也可以表达非反事实; 但是如果加上副词"早"("如果 P,早就 Q")就只能表示反事实。章敏(2016a)探讨了"本来"的反事实语义,指出"本来"有三种可能的语义关系: 正向关系(递进)、零向关系和反向关系(转折)。零向关系中,"本来"需要借助情态动词的帮助,形成"本来+就+情态动词"结构,才能产生反事实语义。并且,不同的情态动词会影响反事实解读。比如,道义情态在这个结构中出现的可能性最大,频率最高,可以产生反事实解读; 而认识情态和动力情态,则不容易产生反事实解读。张莹、陈振宇(2020)则将条件句的标记分为三类: 反事实标记、非事实标记和中性标记。其中,"要不是"是典型的反事实标记,"只要"是典型的非事实标记,"如果"则是典型的中性标记。除了以上的反事实标记以外,白新杰(2021)指出,话语标记"早知道"也可以作为反事实虚拟标记,并且,"早知道+S"可以构成反事实虚拟句。有的学者还从跨语言的角度,对反事实条件句进行了考察。比如,鞠晨、吉田泰谦、袁毓林(2022)比较了汉语和日语反事实愿望句的表达策略,指出汉日都可以通过相关情感表达传递反事实意义,并且反事实意义都倾向于使用条件句来表达。此外,也有学者考察了汉语方言中反事实思维的表达方式。比如,朱华平(2012)认为安徽庐江话的"照讲"也是反事实传信标记。袁毓林、张琳莉(2018)考察了苏州话中的反事实条件句,并总结出了六种常见的格式: 假设连词+VP、否定性连词+VP、V_1仔 O_1+V_2O_2、副词+VP、否定词+VP、VP_1+末+VP_2。张恒君(2019)指出河南孟州方言的"忘了"具有表达反事实意义的作用,并对"忘了"的句法分布、对句子所在主语的人称限制,以及由"忘了"引导的反事实句的语用功能进行了考察。

就汉语普通话而言,有的句式虽然可以表达反事实,但也可以表达非事实。它们在表达反事实时,我们往往会结合一些其他因素来进行判断。比如,Feng & Yi(2006: 1283)在对汉语反事实标记的心理学实验研究中证明,含有"要不是"而被母语者认定为反事实条件句的比例高达 91%,仅次于"原来应该"(转引自章敏 2016b)。也就是说,即使是大家公认的反事实标记"要不是"也不能完全保证该标记所在的条件句一定为反事实条件句。

因此,有的学者对汉语中是否存在反事实标记提出了质疑。比如,蒋严(2000: 264-277)认为汉语没有统一的或强制性的反事实标记,反事实更多的是一个语用解释问题,通过上下文、语境以及日常知识来判断。雍茜(2015)认为汉语中没有形成成熟的 CF 标记(Counterfactual Maker,反事实标记),汉语中的违实句主要通过 HE 标记(Hypothetical Enhancing Marker,假设增强标记)来实现。她指出,违实句是指传达与主观事实相反意义的语句,只有说话者能够确定句子的违实性,听话者需要通过各种词汇语法特征或词汇线索推测出违实义。这些语法特征或词汇线索的使用具有增强句子假设度的功能,故能引发高假设度的违实解,这种标记被称为 HE 标记。HE 标记通过贡献自身语义增强句子的假设度,当假设度增高到一定程度才可以激发违实义。类似于其他 HE 标记语言,汉语违实义的表达没有固定的语法词汇形式,而且部分依赖于语用。

那么真正影响我们将条件句解读为反事实的因素是什么呢? 张莹、陈振宇(2020)对小说文本以及科技文本中的条件句进行了统计,认为汉语条件句的事实性是一个有梯度、有序列的

连续统。根据假设连接词所在条件句的反事实比例，他们把假设连接词大致分为三类，发现它们大致呈现连续统的分布。此外，他们还总结了一些影响反事实倾向的语用因素，包括：时间制约、情感制约、频率制约、数量制约和人称制约。具体结果如下图所示：

		反事实	中性	非事实
连接词	小说文本	要不是/早知道>纵使>假如>要是>就算>假若>	如果>若是>假使>倘若>	就是>哪怕>即使>纵然>只要>万一/一旦/只有/倘使
	科技公文类文本	要不是>假如>假若>倘若>要是>若是>	如果>	只有/只要/即使/一旦/哪怕/万一/就是
特征标记	小说文本	就好了$_{后件}$>早 X$_{前件}$>不是>早 X$_{后件}$>时间词>没有>我/我们>了$_{后件}$>第三人称>	反问$_{后件}$>真/真的>	不>你/你们/您>了$_{2前件}$
	科技公文类文本	就好了$_{后件}$>时间词>早 X>不是>反问$_{后件}$>没有>真/真的>我/我们>了$_{2后件}$>第三人称>	你/你们/您>	不>了$_{2前件}$

表 1　连接词与特征标记梯度（转引自张莹、陈振宇 2020）

张莹、陈振宇（2020）所总结的影响因素实际跟雍茜（2015）所提出的 HE 标记相似。雍茜（2015）认为，汉语中的 HE 标记包括时制特征（主要是过去时）以及语气范畴（汉语中的语气助词“了”）、否定、客观与临近度等；这些因素都可以增强违实表达效果，而不能排他性地标记违实因素。但是，如果同一个句子中拥有两个或两个以上的 HE 标记，那么将这个句子视为违实句的几率会大大增加。根据她的调查，在条件框架下，否定和第一人称同时出现时，违实义的引发概率几近 100%。章敏（2016a）主要对反事实句中的情态进行了分析，指出在不同的反事实句中，情态动词的不同小类可能影响反事实语义的解读以及反事实语义的强度。比如，认识情态动词最容易与反事实共现，同时所参与构建的反事实语义强度高，不易消除；而动力情态和道义情态所参与构建的反事实语义强度低，且可以通过语法手段消除。此外，即使是与反事实语义最容易共现的认识情态类型，也可能会由于其内部可能性的高低对反事实语义的解读造成影响。鞠晨、袁毓林（2021）通过对真实文本语料的考察，指出感叹副词能推动或加强愿望句的反事实意义，通过削弱句子的信息，从而使听话人通过语用推理得到反事实解读。

综上所述，我们在观察反事实条件句的时候，不仅要考虑到有关连词对事实性的影响，而且要考虑其他语用因素（比如，时间、情感、频率、人称等）对于人们识别反事实的作用。

二、法律话语中的事实性构式

事实性表达作为一种语言推理的导航机制，对话语理解、信息抽取、论辩挖掘等研究和实践领域都有着重要作用。特别是法律领域的研究和实践表明，事实性表达可以指引人们对话语中有关事件或事实之间的关联程度做出判断；事实关联程度的不同，可以反映事件或事实之间的因果关联的强弱。在法律领域，这就为法律因果关系判断提供了重要的基础。法律领域的因果关系（causation）有着非常重要的作用，这是认定行为人是否需要承担法律责任的一个

重要考虑因素。但不同事物、行为或事件之间是否存在真正的因果关系，往往难以判断。法学界对原因和结果之间的链接(link)关系有很多讨论，出现了条件说、原因说、相当因果关系说、合法则的条件说、客观规则论等不同的因果关系判断理论。法律工作者试图通过自然法则、经验法则或社会法则等规则，来判断或建构不同行为之间的因果关联。但无论是哪一种理论，法律领域的因果关系判断在本质上都是对事件的事实性的探寻，都是对下面这些核心问题的回答：不同事件之间的事实性关联构成了何种程度的因果关联？是否形成了具有法律意义的因果关系？由于事实性构式指引不同事件之间事实性关联的判断，即不同事件在因果关系上的关联程度；而事实因果关联又影响法律因果关系，进而影响法律审判结果；因而事实性构式在法律领域经常被使用，说话人力图通过不同的事实性构式的语义特性，来表达己方对不同事件之间的事实性关联的意见，并影响听者的语义推理，竭力实现己方的案件事实主张。比如下例：

(1) **辩方**：通过视频和法庭调查可以看到，这个案件实际是王国立在倒卖卡，自己行使违法行为，自己无事生非，自己攻击别人。由于自己的身体原因，我们不能说他是气死的，而是由于他自己的身体原因造成的。一直都是他自己的原因。他自己的原因却让别人承担责任，我认为是不公平的。鉴定中的1%—20%只是说明冲突这一事件对血压升高存在1%—20%的诱因关系，但不能说是直接的因果关系。身体不舒服了，如果他能早点就医检查，可能就不会发生死亡后果了。从视频可以看到，中间两个小时的时间他一直到处转悠，还去买了烟。说明他自己都认为没有什么问题。这个更是和别人毫无关系。别人没法预料到，也不可能存在过错。如果及时就医可能就不会发生后面的后果。没有及时就医是他更重要的过错。

上例来自一个排队纠纷案，当事人王国立、牛筝在排队过程中发生言语冲突和肢体接触，后来当事人王国立死亡。该当事人被证实患有高血压，发生纠纷到死亡期间并无就医的想法或行动。王国立的家属提起控诉，指控另一当事人牛筝和王国立的冲突引起王国立血压升高、身体不舒服，最终导致了王国立的死亡。即控方建立三个事件之间的链式因果关联：冲突→血压升高、身体不舒服→死亡。这个指控实质上就是把“冲突”行为和“死亡”结果进行了因果关联。

在论辩过程中，辩方巧妙地利用了反事实表达(见例中划线部分)建立新的事件关联，并打断了控方提出的因果关联。当事人王国立在“身体不舒服”之后，并没有实施“就医”这一行为。辩方以这一事实为基础进行反事实表达——“如果他能早点就医检查，可能就不会发生死亡后果了”“如果及时就医可能就不会发生后面的后果”，通过反事实表达将“及时就医”行为和“不会死亡”后果进行事实性关联，并引导听众做出新的事实推理，形成新的因果关联：冲突→血压升高、身体不舒服→没有及时就医→死亡。

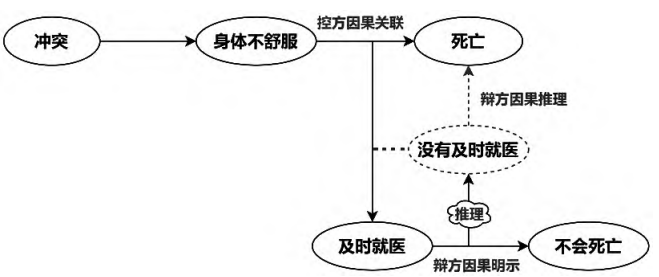


图1　控辩双方的因果关联

从上图可看出，控辩双方对于“死亡”这一结果关联了不同的原因，实现不同的责任归因。辩方更是在控方的责任归因主张的基础之上，利用反事实表达成功地构建了新的责任归因。这一成功的根本在于反事实表达在事实性推理上的特殊导航作用。

反事实表达可形式化表示为“如果 P，则 Q”，其内在的本质就是一种反事实推理。反事实推理是因果推理的重要组成部分。反事实推理可以推断在事件 P 已发生或没有发生的情况下，事件 Q 会发生或者不会发生的概率；反事实表达作为反事实推理的一种形式，反映了事件 P 和事件 Q 之间因果关联的必要性，这就是一种归因的过程。但因果关系并不是一种实质蕴涵（material implication）。“一般认为，引起一定现象发生的现象是原因，被一定现象引起的现象是结果。这种现象与现象之间的引起与被引起的联系，就是因果关系。”（张明楷 2006：61–62）但是对于事件 Y 而言，事件 X 引起了 Y，但事件 X 也只是结果 Y 的一个可能原因。反事实表达将两个事件进行了因果关联，但事件因果关联的远近具有差别。上例中的辩方正是充分考虑到了“死亡”这一结果有多种可能的原因，针对控方提出的“冲突→血压升高、身体不舒服→死亡”这一因果链条，利用反事实表达建构了一个虚拟场景，把“死亡”结果和更近的一个原因“没有及时就医”进行关联，建立了新的因果关联，从而实现了新的责任归因。因果关系具有可撤销和非单调的性质，因此在进行因果推理时，当产生新的原因推理时，旧的原因推理可能会被抛弃。辩方构建的新的责任归因，由于提出了距离“死亡”结果更近的一个原因“没有及时就医”，这就比远因“身体不舒服”更为可靠，关联度更高；因此，对控方提出的责任归因主张造成了直接的对抗和否定。

反事实表达是以现实情况为基础的事实性构式，假设条件句则是以不确定的现实情况为基础的事实性构式，该构式类型在法庭论辩中同样发挥了事实推理的导航作用。例如：

（2）**原告**：<u>如果她没有还手的话，我怎么会受伤呢</u>？

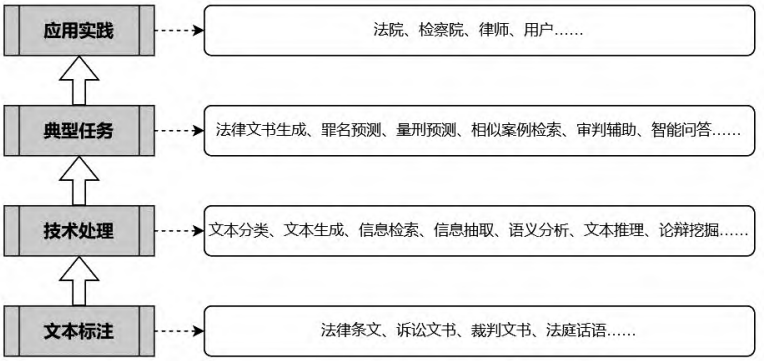
上例来自一个因琐事引起殴打的纠纷案，在殴打过程中被告是否还手打了原告是一个有争议的、有待确定的事件。被告明确否认，原告则以这一有待确定的事实为基础，假设这一有待确定的事实已经发生，但在后件的因果关联中却表示，这一假设和已确定事实发生矛盾。崔玉珍（2021）提到“在这种表达中，条件从句的争议事实起到了对立回声的作用，说话人意图对反方态度或观点发表否定或反对意见，但在从句暂时接受与己方辩护相矛盾的反方陈述，然后在主句显示其推断结果，表明该结果与其他已被接受或承认的事实不一致，从而推断出从句所表达的反方陈述并非为真，实现了分离回声的论证功能。”

可见，在法庭话语中，事实性构式直接影响法庭的事实推理和解释，对法庭话语理解、信息抽取、论辩挖掘等深层应用领域具有重要作用。

三、面向事实性构式的专题性法律语料库

随着人工智能的快速发展，法律领域的智能化研究得到广泛关注和探索。深度学习技术的应用、新模型和新算法的有效迁移，都使得法律智能研究不断深入，推动了从法律实务信息化走向法律智能化。国务院《新一代人工智能发展规划》中明确提出“建设集审判、人员、数据应用、司法公开和动态监控于一体的智慧法庭数据平台，促进人工智能在证据收集、案例分析、法律文件阅读与分析中的应用，实现法院审判体系和审判能力智能化。”从中可以看出，法

律智能是以服务司法实践为出发点、以智能化为目标、具有多重任务设定的人工智能技术应用。我们认为，这种应用应该涉及四个层面：文本标注→技术处理→典型任务→应用实践，不同层面包含不同的内容。具体情况可以图示如下：



当前的法律智能研究有两大特点：

1）法律智能的处理对象多为法律条文、诉讼文书、裁判文书等较为规范、规则的法律文本；

2）法律智能的常用技术目前多集中在文本分类、文本生成、信息抽取和语义匹配等方面。因此，面向的任务场景就集中在法律文书生成、罪名预测、量刑预测和相似案例检索这四个方面。目前的法律智能研究已经进行了很多有益的探索，取得了不少成果。但同时也面临很多的挑战，比如：

1）开放的、动态的、真实情景中的法律语料较为缺乏，因而难以进行真实互动中的法律信息刻画，也就难以进行人类推理和解释机制的建模。

2）高质量标注的训练数据较为缺乏，导致了触及智能核心的法律推理智能研究的探索不够充分。此前的研究，大多是对法律语料进行命名实体的识别。命名实体识别会加入初步的法律领域特性考量，在此基础上增加案例关键词的标注或抽取，然后进行计算和机器学习。这种技术路径在场景相对封闭的数据应用中，可以取得不错的效果。但这些数据的简单标注或处理难以揭示事物之间复杂的语义关系和因果关联，在机器学习中不同层次的要素就无法进行差别识别，因而难以进行有效的法律推理和论辩挖掘。

由上可知，法律智能的深入和推进离不开知识的表示和推理。事实性构式的法律专题性语料库，正是一种面向知识表示和推理的特定语料库；可以标注法律推理过程中已发生的事件、没发生的事件和不同概率事件之间的因果关联，从而可以帮助对事实性表达和事实论证进行结构化分析，最终可以为深度学习知识表示和推理应用提供高质、有效的训练语料。

四、面向事实性构式的专题语料库的构建

下文讨论怎样对法律领域的事实性构式进行标注，标注其中的事实、非事实和反事实三种不同的事实性信息，从而完成面向事实性构式的专题语料库的构建。语料库的构建过程主要包括语料选取及处理、标注内容、标注规范调整以及标注一致性控制四个方面。下面分别讨论。

4.1 语料的选择

本文的语料来源主要来自《薄熙来庭审实录》(下文简称“薄案庭审”),共139,585字。2013年7月25日,原重庆市委书记、中央政治局委员薄熙来涉嫌受贿、贪污、滥用职权犯罪一案,由山东省济南市人民检察院向济南市中级人民法院提起公诉。2013年8月22日,济南市中级人民法院一审公开开庭审理被告人薄熙来受贿、贪污、滥用职权案。《薄熙来庭审实录》根据一审公开开庭的审理转写而成。薄案庭审是我国1996年刑诉法修改确立“控辩式”审判模式以来庭审的经典之作。在庭审过程中,控、辩双方围绕争议的焦点事实,展开了积极、激烈的攻防对抗。其中,事实性构式是双方因果推理的重要语言形式指标。因此,我们尝试选择薄案庭审实录进行事实性构式的标注。

4.2 标注内容

本文的语料库标注聚焦于事实性构式的三种最重要的信息:事实、非事实和反事实,即将事实性信息分为三大类:说话人认为已经发生的事件为事实信息,说话人不确定或不关注事件是否已经发生的为非事实信息,说话人表明事件一定不发生,跟已经发生的事情相反的为反事实信息[①]。例如:

(3) 张三说李四已经走了$_{事实}$。

(4) 如果你来$_{非事实}$,我就走。

(5) 早知道下雨$_{反事实}$,我们就不出来了。

在三类事实性信息的基础之上,我们进一步从三个层次挖掘跟事实性信息相关的表达特征,并进行标注:

1) 构式级别特征

构式级别特征是指事实性表达构式,主要集中在条件句方面,根据事实性信息的三种类型可把条件句分成事实条件句、非事实条件句和反事实条件句。如例(6)-(8):

(6) 既然政府养不起这个公司,还不如撤销掉。

(7) 如果你对这部分事实没有异议,也可以不陈述。

(8) 如果是为了个人,我是完全不会同意的。

2) 语段级别特征

语段级别特征是指事实性表达构式前件和后件所呈现的事实性表达因素。事实性构式主要体现在条件句上,而条件句由前件、后件构成。由于事实性立场及句式结构会影响整个事实性构式的语义推理,因而需要对此进行标注。法庭话语作为一种论辩话语,控辩双方都会根据己方的意图进行案件事实的构建和博弈。并且,不同立场的主体对事件的事实性状态及不同事件之间的关联权重可能会出现分歧。因此,我们需要对事实性构式中的事实性立场进行细化标注。本文将根据法庭角色,区分三种事实性立场:控方事实、辩方事实、被告人事实。而条件句后件的句式结构则会影响事实性构式的推理模式:当后件为陈述句,前件和后件直接建立因果关系;当后件为反问句,需先对反问进行推理得到事件的陈述,然后再和前件进行因果关联。例如:

(9) 公诉人:如果没有薄熙来的安排,薄谷开来和王正刚怎么能将500万进入薄谷

开来实际控制的账户?

(10) 被告人:我对王正刚送来的500万不闻不问不嘱咐就收了,这完全不合情理。

上面两例分别是控方公诉人和被告人主张的事实,都是集中在被告人是否接受了500万贿赂这一争议事实上。在例(9)中,控方建立了“没有薄熙来的安排,则薄谷开来和王正刚不能将500万进入薄谷开来实际控制的账户”这一因果关联。其替代性推理为“有了薄熙来的安排,薄谷开来和王正刚就能将500万进入薄谷开来实际控制的账户”。这一事实主张强调银行账户金额的变动这一具有客观证据的事实,同时通过反事实表达赋予其直接原因为“有了薄熙来的安排”。而在例(10)中,被告人淡化了银行账户金额的变动这一事实,而是强调银行账户金额变动之前的行为。而这正是控方难以提供客观证据的地方。因此,被告人提出了新的因果关联,同时也对控方建立的因果关联进行回击和反驳。

3) 词汇级别特征

词汇级别特征是指事实性构式中影响事实性判断的词汇因素,主要体现在叙实性动词、名词、副词或特定连词、副词上。比如,“怀疑”“隐瞒”“事实”“真的”等叙实性词汇就表明了事件源对事件发生的立场和态度,连词“既然”则通常表明后面从句中的事件是已然发生的状态。

4.3 标注体系和标注规范

语料库的标注体系决定对语料加工的具体方面和对语料的加工程度。如果类别划分过粗,就不能全面、细致地描述语言的复杂现象;如果类别过细,标注信息过于庞大,不仅会增加标注难度,并且关系之间只有细微差异的情况也会使标注结果呈现严重的不一致性。

我们分三个阶段对薄案庭审中的事实性信息进行标注:1) 预标注阶段,该阶段共有三位标注人员对相同的庭审语料进行标注,记录存在疑问的情况;2) 标注规范调整阶段,该阶段对标注一致性及标注规范科学性进行分析和讨论,先探讨标注不一致的情况,接着讨论标注规范需调整或细化的地方,最终形成完善后的新标注规范;3) 正式标注阶段,该阶段标注采取叠加标注策略,相同语料由两人重复进行标注,完成标注后,统计标注不一致的情况,然后讨论并修改标注语料,最终形成面向事实性的专题语料库。

结合前文中分析的三个层级的事实性信息,我们将最终的标注体系总结如下:

类别	符号	说明
事实	FacSen	事实句
	FacPr	控方主张的事实
	FacDe	辩方主张的事实
	FacDef	被告人主张的事实
	FacV	叙实动词
	FacN	叙实名词
	FacAdv	叙实副词
非事实	nonSen	非事实句

续表

类别	符号	说明
反事实	CouSen	反事实句
	CouPrAn	控方主张的反事实前件
	CouPrCo	控方主张的反事实后件
	CouDeAn	辩方主张的反事实前件
	CouDeCo	辩方主张的反事实后件
	CouDefAn	被告人主张的反事实前件
	CouDefCo	被告人主张的反事实后件
	De	反事实后件为陈述句
	In	反事实后件为疑问句
	CouConj	反事实连词
	CouAdv	反事实副词

表 2　标注体系中各变量的说明

为了方便提取语料中的事实性表达，我们采用了 BMES 四元序列标注法，对原始语料进行了标注。其中，B 表示一个词的词首位置，M 表示一个词的中间位置，E 表示一个词的末尾位置，S 表示一个单独的字词，O 表示无关标记。根据以上标签，我们可以将例(8)和例(11)标注如下：

(8) 如果是为了个人，我是完全不会同意的。

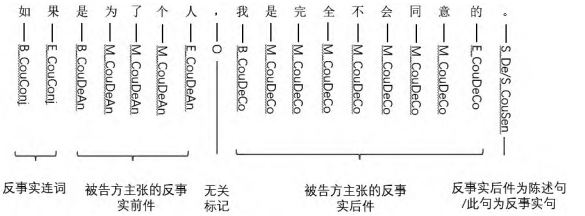


图 2　事实性语料标注样例

(11) 公诉人首先对受贿犯罪接受唐肖林 110 余万元的事实综合答辩。

图 3　事实性语料标注样例

4.4 纠错机制

标注体系和标注规范用以保证语料在录入和标注时的准确率和一致性，而纠错机制是在语料标注完成后统一进行语料标注的正确性和一致性检查。为了统一不同标注者对某些常见情况的标注标准，我们采用了部分的交叉标注，以保证语料标注的正确性。在一致性检查方面，则是采用人工修正的方法。

五、面向事实性构式的专题语料库的应用实例

近年来，随着以裁判文书为代表的司法大数据不断公开，以及自然语言处理技术的不断突破，如何将人工智能技术应用在司法领域，辅助司法工作者提升案件处理的效率和公正性，逐渐成为法律智能的热点。为了促进智能技术赋能司法，中国法律智能技术评测 CAIL(Challenge of AI in Law) 提供了大量司法文书数据作为数据集，并设置了论辩挖掘、论辩理解、信息抽取等多项任务。其中，论辩理解旨在自动识别庭审笔录中辩诉双方的争议观点，并提取案件的争议焦点。而当前争议焦点的提取往往依靠法官人工阅读、整理、分析和归纳，耗费了大量的审判资源。面向事实性构式语料库的建设，尤其是反事实表达的使用，表明了论辩双方在某一事实认定上存在矛盾。这有助于我们提取案件的争议焦点，也可以提高法律工作者的工作效率。在《薄熙来庭审实录》中，被告人大量使用反事实表达来说明自己所认定的事实与公诉人认定的事实存在矛盾。在该案庭审中，有一个情节是被告人薄熙来被指控贪污了 500 万元，我们以“500 万”为关键信息对控辩双方的庭审博弈进行了抽取，并部分摘录如下:

(12) **起诉书指控**: 被告人薄熙来犯贪污罪……2002 年 3 月工程完工后，该单位通知王正刚，拨款人民币 500 万元给大连市人民政府。王正刚遂向已调任辽宁省人民政府省长的薄熙来请示如何处理该款项，薄熙来表示考虑一下再说。一周后，王正刚再次向薄熙来汇报，提出该 500 万元在大连市没有其他人知道，因此提议将该款留给薄熙来补贴家用。薄熙来当即将此事电话告知薄谷开来，并让王正刚跟薄谷开来商议处理。几天后，王正刚到沈阳市薄熙来家中，与薄谷开来议定将该 500 万元转至薄谷开来指定的北京市昂道律师事务所主任赵东平处，供薄熙来家庭使用。**(事实表达)**

被告人: 他在之前他从来没有给我送过钱，也从来没听过我收过钱，在这种情况下他怎么能够断然把这 500 万元给了我，就提出这个建议呢? **(事实表达)**

被告人: 他让我收钱，总得说出一些理由来，他(王正刚) 刚才讲因为我老婆孩子在外面，他想给我补贴一点家用，这个理由对一个领导干部来说，能是一个打动他人心的理由吗? **(事实表达)**

被告人: 谷开来的收入情况非常好，谷开来证实她共办有 5 个分律师所，经济情况非常好。还有，谷开来还给我说瓜瓜也很优秀，有奖学金，我有什么理由担心他们有什么困难呢? **(事实表达)**

被告人: 王正刚说他给我送钱的理由，如果说我收 500 万元，我总得思考、策

划，对于一个贪污犯来说他总得想想这笔钱还有谁知道吧？收这笔钱安全不安全吧？（**反事实表达**）

被告人：按王正刚的说法，是我给薄谷开来打电话时，他在场，在电话里我讲，这是上级某工程的钱，500 万明确提到了，这个事是否合情合理？（**反事实表达**）

被告人：熟悉我的人都知道，和我说话时我先要求他们关手机，我还是一个比较谨慎的人。我会不会通过有线无线，指示谷开来把这 500 万拿下来？（**事实表达**）

公诉人：如果没有被告人薄熙来的同意，王正刚怎么能去找薄谷开来？（**反事实表达**）

公诉人：如果没有薄熙来的安排，薄谷开来和王正刚怎么能将 500 万进入薄谷开来实际控制的账户？（**反事实表达**）

上面的摘录按照庭审的时间先后顺序进行了列举。控方公诉人首先通过事实表达做出己方事实主张，并建立起贪污指控的因果链条；被告人则同时采用事实表达和反事实表达来对控方的部分事实主张进行反驳。控方在对被告人的反驳进行反-反驳时，主要集中在己方具有客观证据而且被告人此前未进行反驳的两点事实主张上“王正刚和薄谷开来商议”和“500 万转至薄谷开来控制的银行账户”，采用反事实表达对被告人之前的反驳进行分离回声式的对抗。控辩双方的论辩图示如下：

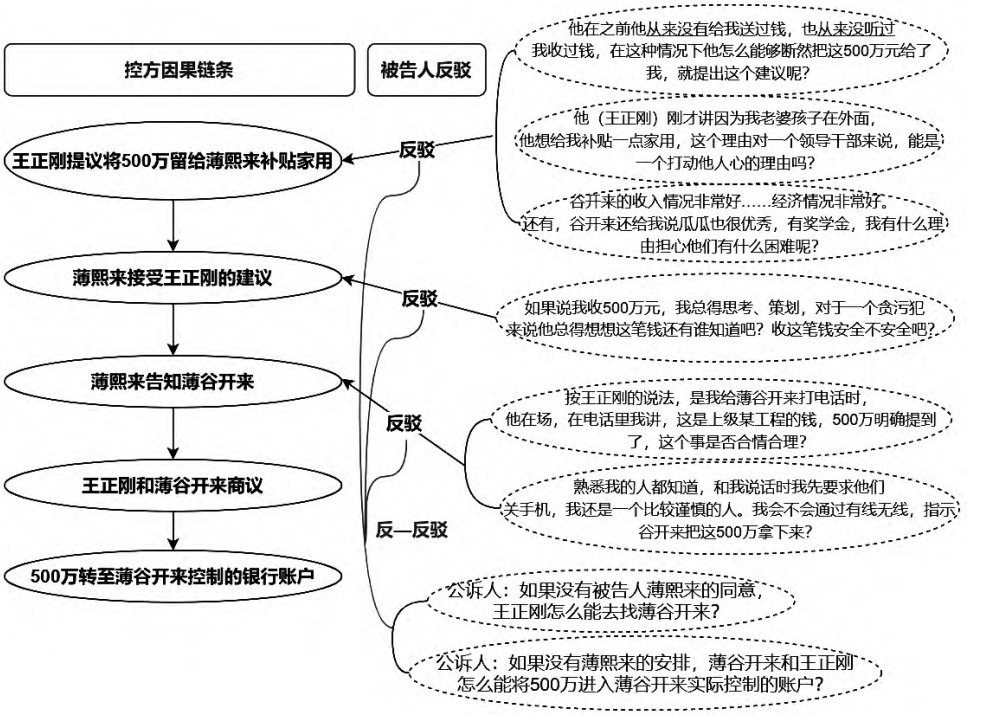


图 4　法庭论辩图示

从上图可看出，利用法庭论辩中的事实性表述，有助于我们挖掘控辩双方的争议焦点，并对双方的事实主张、反驳过程进行快速抽取。值得注意的是，控辩双方在论辩过程中使用了大量的反事

实条件句来表明己方的事实主张。但是，目前的人工智能缺乏对因果关系的理解，不能很好地处理反事实问题（铂尔、麦肯齐 2019: XXIII）。简单来说，就是不能理解反事实条件句所表达的内容。为了让计算机能更好地处理这类问题，我们需要利用前文的标注，将反事实条件句转化为计算机可以理解的形式。例如：

（13）如果没有被告人薄熙来的同意，王正刚怎么能去找薄谷开来？

例（13）是一个反事实条件句，需要利用之前的标注先将后件转换为陈述句，再对其前件和后件分别进行取反的操作，最后将其转化为计算机可以理解的形式。具体结果如下：

（13a）有了被告人薄熙来的同意，王正刚才能去找薄谷开来。

除了反事实条件句，反问句也是计算机处理的难点。我们也需要将其转换为相应的陈述形式。例如：

（14）我有什么理由担心他们有什么困难呢？

例（14）是一个反问句，我们也需要将其转化为陈述句“我没有理由担心他们有什么困难”。

我们运用以上方式，对控辩双方的事实主张进行处理，并提取了具体的事实内容。最终的结果，如下图 5 所示：

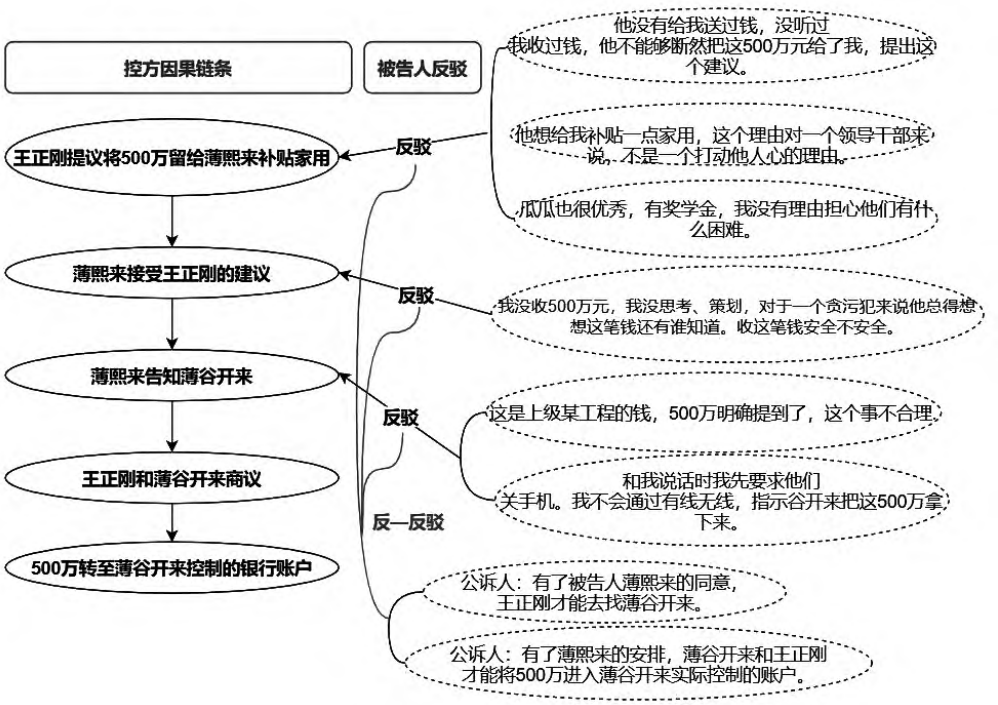


图 5　处理后的法庭论辩图示

相较于图 4，经过处理后的论辩过程更加简洁，并且直接表明了控辩双方的事实主张。这有助于实现计算机自动处理，进而形成结构化表达。并且，法庭事实性表达的结构化可进一步结合深度学习的技术，从而可系统表示、推断和分析法庭论辩过程中出现的不同事实主张或假设、因果链条和可替代性解释；然后通过评估事实主张或假设之间的相关性和权重，来确证或否定事实主张或假设，做出更为合理的事实推论，最终实现事实论证的结构化分析。

注　释

① 本文对事实性信息的分类，主要参考张莹、陈振宇(2020)。

参考文献

白新杰　2021　话语标记“早知道”的反事实与反预期——兼论普通话“早知道+S”的反事实虚拟句，《语言与翻译》第1期。

崔玉珍　2021　法庭反事实表达的论辩研究，《中国语文》第6期。

蒋　严　2000　汉语条件句的违实解释，《语法研究和探索》(十)，商务印书馆。

鞠　晨、吉田泰谦、袁毓林　2022　汉、日语中愿望的事实性及其表达特点，《外语教学与研究》第4期。

鞠　晨、袁毓林　2021　感叹副词对愿望句反事实意义的推动作用，《世界汉语教学》第4期。

李思玮　2019　“如果P，早Q”反事实句研究，华中师范大学硕士学位论文。

雍　茜　2015　违实句的形态类型及汉语违实句，《外国语》第1期。

袁毓林　2015　汉语反事实表达及其思维特点，《中国社会科学》第8期。

袁毓林　2020　叙实性和事实性：语言推理的两种导航机制，《语文研究》第1期。

袁毓林、张琳莉　2018　苏州话反事实条件句的句法形式，《常熟理工学院学报》第3期。

张恒君　2019　河南孟州方言的反事实虚拟句“忘了+S”，《汉语学报》第2期。

章　敏　2016a　“本来”反事实句与情态共现问题研究，《新疆大学学报》(哲学·人文社会科学版)第1期。

章　敏　2016b　“要不是”反事实条件句的情态问题研究，《中南大学学报》(社会科学版)第2期。

张明楷　2006　《刑法学》，北京大学出版社。

张　莹、陈振宇　2020　汉语的反事实条件句与非事实条件句，《汉语学报》第3期。

朱华平　2012　安徽庐江话的反事实传信标记“照讲”，《方言语法论丛》第6辑。

珀尔(Pearl，J)、麦肯齐(Mackenzie，D)著　2019　《为什么：关于因果关系的新科学》，江生、于华译，中信出版集团股份有限公司。

Feng，G. & Yi，L. 2006 What if Chinese had linguistic markers for counterfactual conditionals? Language and thought revisited. Ron Sun & N. Miyake (eds.)，*Proceedings of the* 28*th Annual Conference of the Cognitive Science Society*.

Wang，Yuying 2012 *The Ingredients of Counterfactuality in Mandarin Chinese*. Ph. D Dissertation，The Hong Kong Polytechnic University.

How to Build a Legal-specific Corpus for Facticity Expression Research

Yuan Yulin，Cui Yuzhen，Sun Jing & You Hao

Abstract: This paper discusses how to build a legal-specific corpus for probing facticity expression based on real courtroom trial data. For establishing a complete annotation system and annotation standard，it proposes that the annotation of such a corpus should focus on the following three basic categories of facticity: factual，non-factual and counterfactual，and should specify those categories with formal features from three levels: construction，paragraph and vocabulary. By adopting the BMES quadratic sequence annotation method，the data is well annotated with facticity information. Finally，the paper demonstrates the further applications of this legal-specific corpus with examples.

Keywords: facticity，specific corpus，legal corpus，counterfactual，argumentative research