

基于汉语框架的语义标注方法

刘 勇 魏光泽

(青岛科技大学信息科学技术学院 青岛 266061)

摘 要 标注语料库中句子的语义信息应该要有一套完整的规范体系。利用汉语框架和框架元素体系进行标注,首先介绍基于汉语框架的语义标注方法的概念,接着从句法功能、短语类型和框架元素标注这 3 个方面对汉语框架语义标注的具体规范进行阐述,最后,通过与其他语义标注方法比较分析,得出框架语义标注的特点。

关键词 框架语义,语义标注,中文分词

中图法分类号 TP391.12 文献标识码 A

Method of Semantic Annotation Based on Chinese Framework

LIU Yong WEI Guang-ze

(School of Information Science & Technology, Qingdao University of Science & Technology, Qingdao 266061, China)

Abstract The semantic information labeled in sentences from true corpus needs one complete standard system. Using Chinese framework and the frame elements in a system, the paper explained the concept of semantic annotation based on the Chinese framework, and formulated some rules in syntactic functions, phrase types and frame elements. Finally, the paper analyzed the characteristic of the method of semantic annotation based on Chinese framework by comparing with the other methods.

Keywords Frame semantics, Semantic annotation, Chinese participle

1 引言

计算机语义分析是 21 世纪以来自然语言处理的研究重点之一,目前在主干语义信息以及浅层语义信息的处理方面,已经积累了较为丰富的成果^[1,2]。然而,多数语义表示方法和处理结果却在追求体系的简洁性和处理的可行性的同时,丢失了语言表达中深层次的、复杂的语义信息,导致研究结果无法达到现实社会的应用需要。菲尔墨(Fillmore)^[3]在其格语法的基础上进一步提出了框架语义理论,对词语、句子乃至篇章的语义内容统一图式化的认知情景——“框架”加以描述,以细化的语义范畴表示更丰富的语义内容,使得语义分析结果更加准确。

本文引进、吸收框架语义理论^[4-6]的研究成果,为满足中文信息处理应用需要,研究汉语框架语义分析问题。一方面,确立汉语词汇的框架语义知识库构建方法、句子的框架语义标注规范;另一方面,以语义标注语料为基础,根据短语类型、句法功能以及短语内、外部其他句法语义特征,归纳汉语句子中框架元素的识别规则,并通过实验验证规则的有效性。

2 几种词汇语义知识库

目前语言信息处理领域使用较多的词汇语义知识库^[7]包括英语的 WordNet、FrameNet、VerbNet 以及英汉双语 HowNet、CCD 等,下面分别给予简要介绍。

(1) WordNet。WordNet 是由美国普林斯顿大学米勒领

导,于 1985 年着手构建的英语词汇语义知识库,其词汇工程的理论基础是心理语言学,其出发点并不是为了计算机处理,而是为了探讨和研究语言产生、理解的心理活动机制。因此,它关注的是语言能力的认知基础,米勒称之为心理词汇学。

(2) FrameNet。FrameNet 是由 Fillmore 主持的一个基于语料库的计算词典编纂工程,完全是语言学家在一套系统的语义学理论指导下的一次工程实践,其出发点即为语言的计算机理解服务的,这样就避免了把不同层面的语义问题或者百科知识杂糅在一起,也排除了一些语义范畴上的不确定性。FrameNet 的理论基础是框架语义学^[8],是 Fillmore 在其格语法基础上进一步提出的解释词语意义的语义学理论。

(3) VerbNet。VerbNet 是由美国宾州大学的 Karin Kipper Schuler 于 2005 年开始构建的一个在线动词词典,它将动词划分为不同的类,一个超级类划分为若干子类,每一个动词类都描述了题元角色、论元选择限制条件和“框架”。VerbNet 中的“框架”是用结构化的方式描述动词类的句法特征和语义结构,包括动词的及物和非及物特征、题元角色的句法实现方式以及事件推理知识。

(4) HowNet。HowNet(知网)是由中国科学院董振东等于 1988 年开始构建的,涉及了英语和汉语两种语言的词汇,主要揭示概念的共性和个性,因此,它自称是一个知识系统,而不是一部语义词典。同时,HowNet 还着力反映概念之间和概念的属性之间的各种关系。在 HowNet 中,每一个事件类都标识有一个角色框架。

刘 勇 博士,副教授,主要研究方向为中文分词、智能家居, E-mail: liliangjie914@126.com; 魏光泽 硕士生,主要研究方向为基于语义的中文分词算法。

(5) CCD。CCD(中文概念词典)是北大计算语言所于2000年9月启动的一个 WordNet 框架下的汉英双语语义知识库,直接为机器翻译、信息提取、信息检索等各项应用服务。该项目着眼于 CCD 与 WordNet 的兼容问题,对于 WordNet 中的每一个英语概念,力争找到大体对应的汉语概念,反之亦然。

3 汉语框架语义标注

3.1 汉语框架语义标注的概念

汉语框架语义标注^[10],即针对一个句子中某个词元和该词元所属的框架,对框架元素所在的成分标记框架元素的名称、短语类型和句法功能。设“w”为待标记的字符串,标注样式如下:

$\langle\langle FE-PT-GF \rangle\rangle\%/\langle\langle OTHER w \rangle\rangle\% \langle tgt=[FRAME] w \rangle\langle\langle FE-PT-GF \rangle\rangle\%/\langle\langle OTHER w \rangle\rangle\%$

例如: $\langle\langle cog-np-subj \text{ 科学家 }_n \text{ 们 }_k \rangle\rangle\langle\langle time-dp-adva-s \text{ 正 }_d \rangle\rangle\langle\langle tgt=[\text{查看}] \text{ 观察 }_v \rangle\rangle\langle\langle phen-np-obj \text{ 核武器 }_n \rangle\rangle w$ 。

3.2 汉语框架语义标注的具体规范

3.2.1 分词和词性标注

分词和词性标注采用“分词 2000”系统^[9-10]自动完成,辅以必要的人工校对。该软件处理准确率高,词性分类体系合理,尤其是对未登录词和专有名词标注效果好。词性标注符号体系采用通用的词性标注方式,在本文中不再一一赘述,如:“v”表示普通动词,“n”表示普通名词。

在本文的举例中,为了集中问题,减少符号干扰,例句中将不再显示分词和词性标记。

3.2.2 短语类型标注体系

设立短语类型标注体系时,本文特别注意体系的简洁性,因为框架语义标注的目的是为归纳框架元素求解规则服务的,如果短语类型体系太过繁琐和细微,不仅不利于语言现象的归纳、总结,反而还会加重标注工作的负担。本文短语类型标注体系包含 12 个标记符号,如表 1 所列。

表 1 汉语框架语义标注中短语类型标记体系

标记符号	标记名称及其举例
np	名词性短语 如:资本主义 漂亮的帽子
vp	动词性短语 如:给他一本书 带领学生
ap	形容词性短语 如:特别安静 明显
dp	副词性短语 如:虚心地 非常 已
pp	介词短语 如:在青岛 被他的老师
bp	区别词性短语 如:大型 中型 小型
tp	时间短语 如:战争初期 周末晚上
sp	处所短语 如:村子里 中国内地
mbar	数词准短语 如:一千三百
mp	数量短语 如:两三天 这群 一次
dj	单句句型 如:(我觉得)天气还很冷
fj	复句句型 如:(我相信)如果他喜欢,就一定能做好

3.2.3 句法功能标注体系

设立句法功能标注体系时,本文主要参照周明和黄昌宁设计的依存语法标记体系^[11],以目标词为中心,针对其直接从属成分,标记从属成分所承担的句法功能。同样,在决定取舍的过程中注意了体系的简洁性,分类比较概况如表 2 所列。同时,由于语言表达中有些比较重要的语义内容隐含在非直接从属成分中,本文增加了 ext(外部主语)、par(插入语)、pared(被插入语)、head(中心语)以及 GF-s(支撑词的从属成

分),下面将一一解释。

表 2 汉语框架语义标注中句法功能标注体系

标记符号	标记名称及其举例
subj	主语 如:〈考察团〉得出结论 〈我〉不了解
ext	外部主语 如:〈父母〉试图理解孩子叛逆的内心。
obj	宾语 如:得出〈结论〉 看〈一看〉
atta	定语 如:〈他们的〉理解 引起〈他的〉注意
adva	状语 如:〈向前〉看 〈非常〉了解
comp	补语 如:考查〈完〉 看得〈我眼花缭乱〉
pivt	兼语 如:诱惑〈顾客〉购买 使〈我〉想起
soc	兼语补语 如:诱惑顾客〈购买〉 使我〈想起〉
va	连谓成分 如:〈躺着〉看书
par	插入语 如:〈好奇怪〉,我竟然看到了星星
pared	被插入语 如:据我了解,〈会议进行的并不顺利〉。
head	中心语 如:看见的〈事情〉
GF-s	支撑词的从属成分 如:〈密码〉要重新编制

(1) 关于外部主语

在框架中充当动作主体的框架元素,如:[行为]框架中的行动者,[感知觉]中的感知者,[知道]中的认知者,等等。在复句或者结构复杂的单句中,其所在成分可能与目标词元没有直接的句法关系,但是,对于语言理解来说,尤其是对于信息检索等应用来说,它们是非常重要的语义信息。因此,本文尽量在标注中把这些成分突出出来,于是规定了“外部主语”这个特殊的句法功能范畴,具体包括如下情况:

目标词做述宾短语的宾语,且句子的主语与目标词存在语义联系,如:

〈母亲〉试图理解孩子叛逆的内在原因。

目标词做兼语结构的第二述语,前面的兼语成分为外部主语,如:

他使〈我〉明白,没人能够帮助我。

目标词所在分句承前省略的主体成分,如:

〈我〉走在大街上,看到一个朋友向我招手。

(2) 关于宾语

有些框架的客体成分常常可以表示为小句宾语,甚至可能有标点符号将目标词所在分句隔开,这时,为了完整理解语义,本文把“宾语”这个句法范畴的界限放宽,只要其实际语义内容是目标词的客体,无论在句法上属不属于目标词的从属成分,都划归为目标词的宾语,其短语类型多为 dj/fj(单句/复句),如:

舒展、优美的乐曲声使人们仿佛看到:〈phen-dj-obj 五星红旗在天安门城楼上冉冉升起〉。

接着便看见〈phen-fj-obj 笑颜逐开的男女老少夹着坐垫、小板凳,把侯树忱团团围起来〉。

(3) 关于插入语与被插入语

虽然插入语在表达中主要是一种语用成分,但它是显现在句子表层的,根据范晓的观点,“插加成分与被插加的成分之间从结构上看也是一种偏正关系。插入语像状语、补语一样,也是一种附加语(偏语、从属语),是对核心语(被插加的部分)进行限制、修饰或补充说明”。为满足信息处理的应用需要,本文不对语用和语义给以明确的界限,只要在语义结构中能够表明的东西,都可以加以表示,使得所表示的内容具有更大的信息量。因此,在语义分析中,不回避插入语的表达。在框架语义标注中,插入语必须在目标词所承担的框架中充当一定的框架元素,或者插入语本身就承担框架,否则不表示。以下两组,插入语相对于目标词不充当框架元素,因此不标注。

a. 表示打招呼或应答时: 梁洁, 我们今晚去逛超市, 你要不要一起?

b. 表示感叹语气时: 哎呀, 我没看见。

在句法功能标注中, 插入语表示为“par”, 被插入语表示为“pared”。在语义上, 插入语既可以充当其他框架的框架元素, 自身也可以承担一定的框架, 下面分别加以说明。

插入语做框架元素, 一般是非核心框架元素, 分 3 种情况:

a. 表示说话者本人对目标词所表达的事件的一种看法和态度, 有的表示说话者对于事件感到在意料之中还是意料之外, 有的表示对情况的推测和估计(看来、看起来、想来、看样子、说不定), 有的表示肯定、强调等特定的口气(毫无疑问、没问题、不用说、不可否定、说真的、说实在的、老实说、不错), 这些插入语的句法功能标注为“par”, 框架元素为事件评价, 标为“eve-desc”。如:

〈好奇怪〉, 他居然这么想。

〈毫无疑问〉, 我们应当根据事实进行推断。

b. 表示目标词所表达的事件的事实根据或来源, 为根据元素, 标为“evid”, 如:

〈据了解〉, 我国律师制度建立和发展大体上可分为 3 个阶段, 第一阶段是在 50 年代。

c. 对句中表达的某事物做注释性的补充或说明, 为了获得更详尽的语义信息, 这种注释性的插入语也标注为框架元素, 元素类型与被注释的成分所表示的框架元素一致, 如:

从观念层次出发, 按照哲学观念的不同, 可以把思维方式分为 3 种类型: 〈主观型、客观型和主观客观统一型〉。

有些目标词元自身会以插入语的形式出现, 这时被插入语就成为框架元素, 这与本文只标注目标词的从属成分的原则有所违背, 但对于语义表示却是需要的, 句法功能标注为“pared”, 如:

据了解, 〈按他原来的计划, 他要编完 12 卷全集中的最后两卷, 他要写完回忆录中 60 年代到 90 年代这一部分, 还要完成文集中译文集的编辑工作〉。

4 汉语框架语义标注与普通语义标注的比较

本文通过与其他语义标注方法进行对比来分析汉语框架语义标注的特点。这里选择 Sinica Treebank 语义标记语料库作为比较对象。Sinica Treebank 是台湾“中央”研究院开发的一个中文句结构树资料库, 在结构树中, 标识了中文语义和语法信息。其语义标注以类似格语法的表达模式为基本架构, 称为讯息为本格语法 (Information-based Case Grammar, ICG), 共有 63 个“语义角色”(thematic role)。之所以选择该语料库作为对比对象, 一方面是因为其语义标注体系基本上能够代表传统的其格语法语义表达模式, 另一方面, 该语料库在网上公布了一部分标注语料, 方便本文直接获取标注结果。

取 Sinica Treebank 的 5 个句子, 尝试用以上框架语义标注体系进行标注, 从而比较这两项研究的语义表示方法。下面各组句子中, a 句为 Sinica Treebank 的标注结果, b 句为基于汉语框架语义标注的标注结果, 为了方便阅读, 将需要对比的语义角色字体加黑。

1a. VP(time: Ndabe: 早上 | Head: VE2 看到 | goal: VP(location: Ncdb: 外面 | Head: VC33: 出 | theme: NP(property):

VH13: 大 | Head: Nab: 太阳)))。

1b. 〈time-tp-adva 早上〉〈tgt=[非自主感知]看到〉〈pcpt-dj-obj 外面出大太阳〉。〈perc ini〉

2a. S(agent: NP(Head: Nhaa: 你) | deixis: Dbab: 来 | Head: VC2: 看看 | goal: NP(quantifier: NP(quantifier: Nep: 这 | quantifier: Neu: 一 | Head: Nab: 篮子) | Head: Nab: 梨))

2b. 〈perc-np-subj 你〉来 〈tgt=[自主感知]看看〉〈phen-np-obj 这一篮子梨〉

3a. VP(theme: VP(Head: VL4: 让 | goal: NP:(Head: Nab: 人) | theme: VP(Head: VC2(Head: VC2: 看 | time: Dd: 一 | Head: VC2: 看))) | Head: V-2: 有 | range: NP(quantifier: Nep: 什么 | Head: Nad: 关系))

3b. 让 〈perc-np-ext 人〉〈tgt=[自主感知]看〉〈dur-action-mp-obj 一看〉有什么关系 〈phen ini〉

4a. VP(location: PP(Head: P19: 从 | DUMMY: NP(property: Naa: 水 | Head: Ncda: 里)) | Head: VE2: 看见 | aspect: Di: 了 | goal: NP(Head: Nhab: 自己))

4b. 〈ground-pp-adva 从水里〉〈tgt=[非自主感知]看见〉〈null 了〉〈pcpt-np-obj 自己〉

5a. S(agent: NP(Head: Nba: 国强) | location: PP(Head: P19: 从 | DUMMY: NP(Head: Nab: 车窗)) | target: PP(Head: P62: 向 | DUMMY: NP(Head: Ncda: 外)) | Head: VC2: 看)

5b. 〈perc-np-subj 国强〉〈loc-perc-pp-adva 从车窗〉〈dir-pp-adva 向外〉〈tgt=[自主感知]看〉

首先, 需要强调的是, Sinica Treebank 和汉语框架语义知识库的标注目的有本质的差别, 前者标注整个句子的句法语义信息, 后者只标注句子中与目标词有关的句法语义信息。因此, 前者在一个句子中表示出了 3 个动词的语义搭配信息(如 3a), 对于每一个短语来说, 则一直分析到最底层, 汉语框架语义标注只表示了一个动词的语义搭配信息, 不表示其他动词的语义信息(如 3a 中“有什么关系”未作表示), 对于框架元素所在成分来说, 也只标注最大结构, 不再细分层次。

本文未对两种标注体系的句法部分进行比较, 只看相对关注的词语, 即“看、看看、看见、看到”, 二者对从属于这些词语的成分所标注的语义角色有什么差异。首先, 比较 1a-1b 和 3a-3b, b 组比 a 组多了一个零形式标记(ini), 1b 表示出了感知着的缺省, 3b 表示出了现象的缺省。让机器真正理解意义, 不能停留在句子表层形式, 而是要处理一些隐含意义, 这样才能向“智能”的方向迈进, 零形式标记虽然还没有直接表示出缺省的成分的具体所指, 但毕竟已经为隐含意义的表示留下了接口, 应该说比 a 组稍微延伸了一步。

比较 1a, 2a-1b, 2b, a 组中“看到”和“看看”后面的宾语语义角色是一样的, 即 goal; 而 b 组对“看到”标注了 pcpt(印象), 对“看看”标注了 phen(现象), 二者的区别已在前文描述, 这里不再赘述。汉语框架语义标注的语义角色更加细化一些, 有了这样的信息, 应用系统才有可能分辨出 1 句和 2 句语义上的差别。

比较有关时间的语义表示, 1a 中“看到”的状语“早上”与 3a 中“看”的准宾语“一看”都表示为“time”, 根据这种语义表示, 分辨不出二者的区别, 相应的 b 组, 前者表示为“time”, 后者表示为“dur-action”(动作时量), 这样一个应用系统才有可能在寻找诸如“什么时间看到的?”的答案时, 把 1 作为候选

项,而在寻找“看了多长时间”的答案时,把₂作为侯选项;如果都标注为 time,就多了一个干扰选项。

比较_{3a}和_{3b},其中“人”在_{3a}中标记为“goal”,这个语义角色是相对于前面的述语动词“让”来说的,是“让”的客体,它并不表示“人”相对于后面的“看”的语义角色;_{3b}在以“看”为目标词标注时,“人”表示为 perc(感知者)。可见, Sinica Treebank 以“显性”语法关系为依据,只有具有直接句法关系的成分之间才表示出语义角色,而汉语框架语义标注体系中,表示出了一部分隐性的语法关系。

比较₄和₅中跟空间有关的语义成分的标注;_b组将表示 location 的“从水里”和“从车窗”通过是否自主感知加以区分,而_a组则视为一致。₄强调通过水看见了自己,感知的是事件发生的外部条件,就是说所看到的事物(自己的影子)所在的场所;₅强调的是“看”,“从车窗”只是感知者“国强”所在的位置,“车窗”不该看做感知对象。因此,比较而言,_b组的分词结果更恰当。设想一个问答系统,在回答人或物体在什么位置这样的问题时,显然_b组的标注会使得系统输出正确答案的可能性更大。

结束语 本文以框架语义理论为基础,构建汉语语义分析的范畴知识体系,其语义角色细化,语义信息丰富,使得计算机语义理解更加精确化,始终采用基于语料库的方法,避免“从意义到意义”的理性思维,使研究结果可靠、实用;用基于规则的方法解决语义标注歧义问题,避免了基于统计的方法面临的语义资源匮乏、数据稀疏的问题,通过实验证明本研究所归纳的规则对于解决语义角色标注问题十分有效。

语义标注因汉语框架语义范畴的知识体系而有众多更加细化的语义角色,其标注方法体系针对汉语特殊句法形式给出了具体的处理方案。因此,与其格语法体系表达模式相比,框架语义标注模式识别出了更多的语义内容,语义描述也更加准确,能够为计算机应用系统提供更加丰富而实用的信息。

参考文献

(上接第 88 页)

- [7] 林萍,黄卫东. 基于 LDA 模型的网络舆情事件话题演化分析[J]. 情报杂志, 2013, 32(12): 26-30
- [8] 王晓峰,张大鹏,王绯,等. 基于语义轨迹的视频事件探测[J]. 计算机学报, 2010, 33(10): 1845-1858
- [9] 刘宗田,黄美丽,周文,等. 面向事件的本体研究[J]. 计算机科学, 2009, 36(11): 189-192, 199
- [10] 张旭洁,刘宗田,刘炜,等. 事件与事件本体模型研究综述[J]. 计算机工程, 2013, 39(9): 303-307
- [11] Shaw R, Troncy R, Hardman L. Lode: Linking open descriptions of events [J]. The Semantic Web, 2009, 31(8): 153-167
- [12] Kaneiwa K, Iwazume M, Fukuda K. An upper ontology for event classifications and relations [J]. AI, Advances in Artificial Intelligence, 2007, 11(8): 394-403
- [13] Corda I, Bennett B, Dimitrova V. A Logical Model of an Event Ontology for Exploring Connections in Historical Domains [C]// Proceedings of the 10th International Semantic Web Conference (ISWC). Bonn, Germany, 2011: 1107-1114
- [14] Liu Wei, Xu Wen-jie, Fu Jian-feng, et al. An extended description logic for event ontology [C]// Proceedings of the 5th International Conference on Grid and Pervasive Computing(LNCS). Springer, 2010: 471-451
- [15] Nilsson J F. A concept object algebra CA [C]// Proceedings of

- [1] Gildea D, Jurafsky D. Automatic Labeling of Semantic Roles [J]. Computational Linguistics, 2002, 28(3): 245-288
- [2] Hellen, Lars, Beermann D. Classification of Prepositions Senses for Deep Grammar Applications [C]// Proceedings of the Second ACLSINSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications. Colchester, 2005: 74-83
- [3] Charles F. Frames and the Semantics of Understanding [C]// Quaderni di Semantica. 1985: 222-254
- [4] Baldwin. Prepositions in Applications: A Survey and Introduction to the Special Issue [J]. Computational Linguistics, 2009, 35(2): 119-149
- [5] Patrick S-D. Syntactic and Semantic Frames in PrepNet [C]// Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJ-CNLP-08). Hyderabad, 2008: 763-768
- [6] Jorgensen F, Lønning J T. A Minimal Recursion Semantic Analysis of Locatives [J]. Computational Linguistics, 2009, 35(2): 229-270
- [7] 陈群秀. 一个现代汉语语义知识库的研究和实现 [C]// 中文信息处理前沿进展, 中国中文信息学会成立二十五周年学术会议论文集, 2006: 172-181
- [8] 冯志伟. 从格语法到框架网络 [J]. 解放军外国语学院学报, 2006, 29(3): 1-9
- [9] 贾君枝, 毛海飞. 汉语框架网络问答系统问句处理研究 [J]. 图书情报工作, 2008, 52(10): 55-57
- [10] 由丽萍, 范开泰, 刘开瑛. 汉语语义分析模型研究述评 [J]. 中文信息学报, 2005(6): 59-65
- [11] 由丽萍. 构建现代汉语框架语义知识库技术研究 [D]. 上海: 上海师范大学, 2006

the 2nd European-Japanese Seminar on Information Modeling and Knowledge Bases, V: principles and formal techniques. 1993: 121-132

- [16] Nilsson J F. A logico-algebraic framework for ontologies [C]// Proceedings of the 1st International OntoQuery Workshop. 2001: 189-110
- [17] Nilsson J F. An algebraic logic for concept structures [C]// Preprint of the 3rd European-Japanese Seminar on Information Modeling and Knowledge Bases, V: principles and formal techniques. 1993: 75-81
- [18] Wille R. Formal concept analysis: mathematical foundations [M]. Springer Verlag, 1999: 117
- [19] 时百胜, 余泓. 概念知识表示和推理 [J]. 小型微型计算机系统, 2006, 27(9): 1618-1622
- [20] 车敦仁, 周立柱. 概念代数——新一代数据库系统的理论 [J]. 计算机研究与发展, 1996, 33(1): 32-38
- [21] Zhang Xu-jie, Liu Zong-tian, Liu Wei, et al. Research on Event-Based Semantic Annotation of Chinese [C]// 2012 IEEE International Conference on Computer Science and Network Technology (ICCSNT). 2012: 1883-1888
- [22] 付剑锋, 刘宗田, 刘炜, 等. 基于层叠条件随机场的事件因果关系抽取 [J]. 模式识别与人工智能, 2011, 24(4): 567-573