面向汉语自动分析的语言特征工程研究

冯志伟[1],程　勇[2]

(1. 杭州师范大学　外国语学院,杭州 311121;2. 鲁东大学　文学院,山东 烟台 264039)

摘　要:语言特征工程研究是实现汉语自动分析的关键所在,其中包括描述单词本身固有特点的"静态特征",以及描述具体句子中单词之间关联的"动态特征"。如何从静态特征自动推导出动态特征是计算机进行自动分析的一个难点。以大规模标注语料库为数据基础,通过对汉语自动分析过程中涉及到的复杂语言特征集进行系统的梳理,总结关于句法成分、语义角色以及词汇语义之间的制约规则,用以指导计算机实现从静态特征到动态特征的自动推导,从而为实现汉语的自动分析提供理论指导和实践支持。

关键词:静态特征;动态特征;词汇语义制约规则;汉语自动分析

中图分类号:H08　　**文献标志码**:A　　**文章编号**:1673-8039(2020)05-0055-05

汉语自动分析的关键在于让计算机能够理解文本句子中所表达的含义,而理解的过程涉及到词法、句法、语义、语用等不同层面的信息处理,其中一个重要问题在于如何从词典中提供的单词的词汇语义特征出发,自动地推导出句子中各个成分的句法功能,进而获得句子的语义角色关系。一直以来,这个问题往往成为计算机进行自然语言处理的一大瓶颈,亟需为构建更加精确、更加细致的相关语言特征来处理和解决这一难题。

一、汉语自动分析的困难

早在 20 世纪 50 年代,冯志伟等学者就开始了对汉语自动分析的探索[1],开始尝试用 Chomsky(乔姆斯基)的短语结构语法(phrase structural grammar,简称 PSG)来进行汉语的自动分析。然而当应用短语结构语法到汉语的句法分析时,会面临句法歧义的问题。比如在汉语中可以说"鸡吃了",实际上有两个意思,一是"鸡吃东西了",即鸡作为施事主语,另一层意思是"鸡被吃了",在这里鸡变成了受事主语。然而在短语结构语法中,鸡都是作为名词短语 NP 成为主语,吃都是作为动词短语 VP 成为谓语,句子的规则都是:S→NP+VP,其中 S(Sentence)表示的意思是句子,可以看到,虽然这个句子有不同的含义,一种是被动句,另外一种是主动句。但如果让计算机使用短语结构语法来对这个句子进行句法分析,最后分析得到的树形图的整个上层都是 S→NP+VP,完全无法区分句子在语义层面的差异。而在中文里这样的句子到处都是,因此,如何根据汉语自身的特点来改进短语结构语法,使其适用于面向汉语的自动分析,成为一个亟需解决的问题。

基于对上述现象的观察,国内外的不少学者提出了一些手段来改进短语结构语法。例如卡普兰和布列斯南在提出的"词汇功能语法"(Lexical Function Grammar,简称 LFG)、马丁·凯依提出的"功能合一语法"(Functional Unification Grammar,简称 FUG)、盖兹达等提出的"广义短语结构语法"(Generalized Phrase Structural Grammar,简称 GPSG)、珀拉德提出的"中心语驱动的短语结构语法"(Head-Driven Phrase Structural Grammar,简称 HPSG),等等,都采用了复杂特征来描述自然语言。在国内,冯志伟等学者同样对该问题进行了深入思考[2],并提出了"多叉多标记树模型"[3](Multiple-labeled and Multiple-branched Tree Model,简称 MMT 模型),该模型最大的特点是使用了"多标记树结构"来代替"单标记树结构"。以

收稿日期:2020-06-05

基金项目:国家社科基金一般项目"基于大规模标注语料库的语义角色句法实现的词汇语义制约研究"(12BYY123)

作者简介:冯志伟(1939—),男,云南昆明人,文学博士,杭州师范大学外国语学院特聘教授;程勇(1987—),男,山东淄博人,文学博士,鲁东大学文学院讲师。

“小王吃了”和“面包吃了”为例,这两个句子在句法结构层面大体相同,然而在句子的词汇语义层面,两者却存在着很大不同,比如说“小王”在语义上的含义是属于“人”这一范畴,而“人”通常是“吃了”这个动作行为的施加者,而“点心”在词汇语义上属于“食品”范畴,因此一般作为“吃了”动作的对象。而在短语结构规则 S→NP+VP 中,MMT 模型把 NP 进一步加以分割,增加了若干特征来进一步描述 NP。比如在“小王吃了”这句话中,NP 被分解为包含“NP ‖ 人”两个特征,而在“面包吃了”中,NP 则包含了“NP ‖ 食品”两个特征,这两个规则前面的“NP”都从语法层面表示其是一个名词短语,后面的“人”与“食品”都表示了相关的词汇语义特征,因此就在计算上把两者区分开来了。总体来说,MMT 模型能够把句子中蕴含的不同性质的信息进行充分表示。

以 MMT 模型为代表的模型被应用到了机器翻译等汉语自动分析任务中,取得了一定的进展。然而,这些模型在运行过程中往往需要用到不同层面的复杂语言特征,比如词类信息、词汇语义特征等静态特征以及句法信息、语义角色标注信息、逻辑语义关系信息等动态特征。其中静态特征是比较容易从词典中获取的,让计算机直接查询便可以获得。然而句子中各个单词之间与各个词组之间的动态特征却很难获得,需要计算机根据已经查出的静态特征,再根据有关的句法语义规则进行复杂的计算才可以获得,其中这些句法语义规则往往是根据研究者自己的语言直觉编写,难免会有主观片面的弊病,使得计算机难以对汉语进行精确全面地自动分析。为此,冯志伟为 MMT 模型设计了汉语复杂特征集(complex feature set)[4]。

二、汉语的复杂特征集

在汉语的自动分析中,要想描述汉语句子,往往不能单单使用诸如词类、词组类型等简单特征来进行描述,往往无法区分各种歧义现象,达不到汉语自动处理的目的。因此需要设计更广泛、更复杂的语言特征,比如词类特征、词组类型特征、句法功能特征、语义角色关系特征、词汇语义特征、逻辑语义关系特征,等等,这些特征共同构成了汉语的复杂特征集,而汉语的复杂特征集可由其特征与对应的特征值来构成。对于汉语的自动分析来说,所涉及的主要语言特征有以下七个方面。

(一)词类特征

词类是用于描述词的语法类别,通常来说,词类主要包含以下类别:名词、动词、形容词、副词、介词、处所词、方位词、数词、量词、体词性代词、谓词性代词、助词、连词、拟声词、时间词、区别词、语气词、感叹词等 20 个词类,当然这些词类还可以进一步地细分,比如形容词词类又可以分成性质形容词和状态形容词两个子类。

(二)词组类型特征

词组类型特征主要用于描述句子中所包含的短语结构,通常包含动词词组 VP、名词词组 NP、形容词词组 AP、数量词组 MP 等共 4 组。

(三)单词语法特征

孤立的单词往往也具有语法特征。以数词和量词为例,“狗”与“牛”虽都为名词,但前面可接的量词就截然不同;而对于动词来说,及物性、不及物性则可以看作是其固有的语法特征;另外动词的“价”(valence)是动词的另一个固有语法特征,即动词对于其前后相关词语的语法约束,因此也看成是动词的固有语法特征。单词的固有语法特征记为 GRM。其中包括(1)动词及物性:“及物”和“不及物”;(2)动词价:一价(咳嗽)、二价(吃)、三价(给)。

(四)词汇语义特征

单词的词汇语义特征指的是词汇的语义类别,它是单词固有的语义特征,而不是具体句子中单词之间的语义角色关系。单词词汇语义特征记为 SEM,包括:物象、物资、现象、时空、揣度、抽象、属性、行动等类别及相关的值。这些词汇语义特征都可以在语义词典中进行标注,成为单词本身固有的属性。比如《新编同义词词林》就是以词汇语义分类作为基础的。

(五)句法功能特征

现代汉语中的词组类型和句法功能之间的关系错综复杂,往往没有明确的一一对应关系,因此在汉语句子的自动分析中,必须注意句法功能特征,记为 SF,包括:主语、谓语、宾语、定语、状语、补语、述语、中心语等等,其中 SF 的值也可以包含子值,比如直接宾语和间接宾语都可以看作是宾语的两个子值。

(六)语义角色关系特征

与句法功能特征类似,语义角色关系特征往往也不是单词本身固有的,而是需要计算机在进行句法语义分析的过程中动态运算得到。因为只

有在实际的句子中,不同的词之间才能够产生语义角色关系,常见的语义角色关系包括:施事、受事、与事、关涉、时刻、时段、时间起点、时间终点、空间点、空间段、空间起点、空间终点、初态、末态、原因、结果、工具、方式、目的、条件、作用、内容、范围、论题、修饰、比较、伴随、判断、陈述、附加等共计30个值。

(七)逻辑关系特征

汉语的句子可以看成一个逻辑命题,其中的谓词与各个论元之间往往存在着逻辑关系,如下所示:

论元0:主语

论元1:直接宾语

论元2:间接宾语

可以根据论元的情况来检验和判断汉语句子在逻辑关系上是否正确,进而分析得到整个句子的逻辑结构。

上面所列举的各类特征可以分为两类:一类是"静态特征",即是单词本身所固有的语言特征,其中词类特征、词汇语义特征、单词语法特征都可以看作属于这一类,其共同点是可以在词典中独立地把这些特征标记出来。而其余的词组类型特征、句法功能特征、语义角色关系特征、逻辑关系特征等是单词之间发生实际关联时才产生出来的特征,因此叫做"动态特征"。在汉语自动分析中,静态特征是计算机进行运算的基础,而动态特征则在很大程度上决定了一个汉语自动句法分析和语义分析系统的质量高低。因此,如何根据单词的静态信息自动地推导出句子中各个成分之间的句法功能关系、语义角色关系等动态信息,也就是如何从词典中提供的单词的词汇语义特征出发,自动地推导出句子中各个成分的句法功能,进而获得句子的语义角色关系,从而让计算机理解句子所表达的含义,便是汉语自动处理的关键与难点所在,需要对词汇语义特征与句法功能和语义角色之间的制约关系进行充分研究。为此,根据MMT模型的理念,参考MMT模型设计的汉语复杂特征集,鲁东大学基于语料库进行了语义角色句法实现的词汇语义制约研究,这是一项艰巨的语言特征工程(language feature engineering)。

三、语义角色句法实现的词汇语义制约研究

汉语缺乏严格意义的形态标记,词类与句法成分缺乏一一对应的关系,句法语义关系与句法结构并不完全一致,存在着错综复杂的关系。一个词入句后充当何种句法成分、语义角色因句法环境而改变,汉语信息处理依靠词类、句法树、语义角色很难有效地进行句法分析、语义理解。因此必须找到一个初始的特征即词的词汇语义类,建立词的词汇语义类、词类、句法成分、语义角色之间的关系,以词的词汇语义类为抓手,才能有效地解决汉语句法分析、语义推导的问题。关于句法语义关系的研究,国内外一些学者已经开展了一些探索工作[5-9],取得了骄人的成果,但还存在一些不足,比如缺乏基于大规模真实语料的量化研究,这限制了已有工作在应用层面上的价值。

亢世勇团队在对现有工作分析的基础上,尝试进行面向信息处理的、基于大规模真实文本标注语料库的语义角色句法实现的词汇语义制约研究,并且获批国家社科基金项目"基于大规模标注语料库的语义角色句法实现的词汇语义制约研究",经过几年的研究与实践,形成了著作《现代汉语语义角色句法实现的词汇语义制约研究》(以下简称《词汇语义制约研究》),即在构建句法语义标注语料库的基础上,探索哪个词汇语义类的词语充当哪些语义角色时可能映射为哪些句法成分,其中有什么样的规律特征。下面将从基本思路,语料库构建和规律发掘等三个方面对相关工作进行介绍。

(一)基本思路

研究团队通过选取大规模真实文本的现代汉语语料,进行分词、词性、句法成分、语义角色标注,并对涉及到的每个谓语动词、名词进行词汇语义类的标注,建立句法结构、语义结构,词类、句法成分、语义角色、词语词汇语义类对应的语料库,在此基础上进行归纳、统计、分析,总结出系统的语义角色映射为不同句法成分时不同词汇语义类的优先度不等式序列[10-11]。

词汇语义往往可以决定一个论元的语义角色,而语义角色通常可以映射为句法成分,因此词汇语义也就一定程度上预示、决定了句法成分。图1是词汇语义和句法语义关系图。图中句法配位是指语义角色可以映射成为句法成分的关系。句法填位指词汇语义可以投射为句法成分的关系,句法配位和句法填位相互影响,将贯穿整个研究当中。同时,研究者把语义角色与句法成分之间发生关系的手段称为映射,而把词汇语义和句

法成分发生关系的手段称为投射[12-13]。

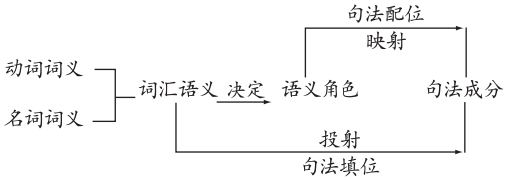

图1　词汇语义与句法语义关系图

语义角色是一个有层级的体系,其中语义角色与目标动词之间的关系亲疏远近并不完全一样,通常语义角色可以分为核心与非核心两种,其中核心语义角色是指那些与动词关系最为密切的角色,比如施事、受事等,一般出现在主语与宾语位置上。而其他的非必须的语义角色则为非核心语义角色。该研究采用的具体语义角色分类如表1所示:

表1　本项目对语义角色的层级分类表

核心语义角色	主体论元	施事、当事、领事、共事
	客体论元	受事、客事、与事、结果、致事、系事、分事
非核心语义角色	凭借论元	工具、材料、方式、原因、目的
	情景论元	时间、处所、方向、范围、同源、数量、基准

(二)语料库建设

该研究在构建大规模真实标注语料库的基础上进行统计研究。语料标注包括词性标注、句法成分标注、语义角色标注、核心词词汇语义类标注。在充分梳理文献,吸收前人研究成果的基础上,确定了词类体系、句法成分、语义角色体系及标记如下。

1. 词类体系及标记采用北京大学计算语言学研究所的词类体系及标记(18类):

名词(n)、动词(v)、形容词(a)、时间词(t)、处所词(s)、方位词(f)、区别词(b)、副词(d)、状态词(z)、代词(r)、数词(m)、量词(q)、叹词(e)、拟声词(o)、介词(p)、连词(c)、助词(u)、语气词(y)。

2. 句法成分及标记采用清华大学周强制定的句法成分及标记(8个):

主语块(S)、述语块(P)、宾语块(O)、定语块(A)、状语块(D)、补语块(C)、兼语块(J)、独立语块(T)。

3. 词汇语义体系及标记均采用梅家驹等主编、上海辞书出版社1983年出版的《同义词词林》的分类体系及标记,共12类:

人(A)、物(B)、时间与空间(C)、抽象事物(D)、特征(E)、动作(F)、心理活动(G)、活动(H)、现象与状态(I)、关联(J)、助语(K)、敬语(L)。每一个大类还可以再分为若干个中类,在大类字母之后再加一个小写字母表示,例如,大类A可以分为泛称(Aa)、男女老少(Ab)、体态(Ac)、籍属(Ad)、职业(Ae)、身份(Af)、状况(Ag)、亲人眷属(Ah)、辈次(Ai)、关系(Aj)、品性(Ak)、才识(Al)、信仰(Am)、丑类(An)等中类。

研究团队构建了《中小学语文课本标注语料库》,其主要来源是相对比较成熟的人教版初中、高中课本,总共约70万字,在该库的基础上,对相应的语料进行了人工标注,标注语料的示例如下:

[S 夏天/t]D1　[P 到/v]V1　了/y　,/w　[S 小树/n]S2　[D 给/p　爷爷/n]T2　[P 撑/v]V2　[C 开/v[O 绿色/n 的/u　小伞/n]O2　。/w　[S 爷爷/n]D　[D 不/d　[P 热/a]V　了/y　。/w

(三)词汇语义制约规则发掘

在标注好的语料的基础上,《词汇语义制约研究》系统、完整地总结了所有语义角色投射为句法成分的词汇语义制约规则,即一系列词汇语义类优先度不等式,比如充当当事主语的义类:A>D>B>E>C>H>I>G>J>K>F;充当当事兼语的义类:A>D>B>E>C>G>H>I;充当当事宾语的义类:D=B>A>C=H;充当当事状语的义类:A>B>D>C>E(注:其中字母即为《同义词词林》中语义大类的标记)。这样的优先度不等式为计算机自动句法语义分析提供了重要的依据,我们可以根据词汇语义类的优先度来优化句法语义分析的规则,从而提高自动句法语义分析的效率[14-15]。

在标注语料库的基础上,《词汇语义制约研究》分析了语义角色句法实现的词汇语义制约的特点。通过对比词汇语义类优先度不等式发现,有些词汇语义类可以充当某些语义角色映射为某种句法成分,而有些则不行。例如,词汇语义特征为A(人)的语义角色充当句法成分的百分比各有不同:

施事语义角色映射为主语的比例为98.78%,映射为状语、宾语的比例为1.22%;

当事语义角色映射为主语的比例为92.07%,映射为兼语、状语、宾语句法成分的比例为7.93%;

与事语义角色映射为状语的比例为54.55%,映射为宾语、补语、兼语的比例为45.45%;映射为宾语、补语、兼语的比例为45.45%;

客事语义角色映射为宾语的比例为62.34%，映射为兼语、主语、状语、补语的为37.66%；

领事语义角色映射为主语的比例为95.78%，映射为兼语、宾语的为4.22%；

受事语义角色映射为宾语的比例为89.85%，映射为主语、状语的为10.15%；

基准语义角色映射为状语的比例为89.91%，映射为状语、宾语的为10.09%；

范围语义角色映射为状语的比例为89.86%，映射为主语的为10.14%；

方向语义角色映射为状语的比例为95.65%，映射为补语的比例为4.35%；

处所语义角色映射为主语的比例为36.84%，映射为状语的31.58%，映射为宾语的为21.05%，映射为补语的为10.53%；

目的语义角色映射为状语的比例为90.91%，映射为主语的为9.09%；

结果语义角色映射为宾语句法成分的比例为80.00%，映射为主语的比例为20.00%。

另外研究团队还发现，在不同的序列中每个词汇语义类所处的地位也是不同的，这反映出该词汇语义类优先度的差异。例如，词汇语义特征为A(人)的单词或词组充当各个语义角色的百分比各不相同：施事(60.86)>当事(21.72)>与事(4.87)>系事(3.88)>客事(3.26)>领事(2.54)>共事(0.81)>受事(0.71)>基准(0.56)>范围(0.35)>方向(0.12)>处所(0.10)>方式(0.06)=目的(0.06)>致事(0.04)>结果(0.03)=原因(0.03)>分事(0.01)=同源(0.01)，词汇语义特征为A(人)的单词或词组不能充当材料、时间、数量等语义角色。

其他词汇语义特征的语义角色在充当句法成分上也有着不同的分布规律，因此可以根据单词词汇语义特征的优先度，来推算它们充当不同语义角色的可能性；也可以根据句子中不同的句法成分来反推出它们可能表示的语义角色关系，进而可以制定一系列的计算机可读规则，为汉语的自动分析提供指导，有着重要的应用价值。

汉语语言特征工程的研究是实现汉语自动分析的关键所在，其中词性、词汇语义等静态特征可以从词典等信息知识库中获得，而句法成分、语义角色等动态则只能由静态特征推导而来。本研究在标注大规模标注语料库的基础上，采用定性与定量分析相结合、形式与意义相结合的方法，系统总结了语义角色映射为不同句法成分时不同词汇语义类的优先度不等式序列。同时，基于大规模语料库中所反映的汉语的真实状况，挖掘词汇语义、句法语义之间的制约规则，深化了现代汉语的语义研究，为汉语自动分析提供可靠的数据，对于进一步提高汉语信息处理水平提供了借鉴思路，具有重要的理论意义和应用价值。

参考文献：

[1]冯志伟.汉语自动分析的MMT模型[EB/OL].(2016-06-23)[2020-02-06].http://blog.sina.com.cn/.

[2]冯志伟.形式语言理论[J].计算机科学,1979(创刊号).

[3]冯志伟.汉语句子的多叉多标记树形图分析法[J].人工智能学报,1983(2).

[4]冯志伟.中文信息MMT模型中多值标记集合的运算方法[J].情报科学,1994(3).

[5]陈昌来.汉语语义结构中工具成分的性质[J].世界汉语教学,1998(6).

[6]陈平.试论汉语中三种句子成分与语义成分的配位原则[J].中国语文,1994(3).

[7]储泽祥,彭建平.处所角色宾语及其属性标记的隐现情况[J].语言研究,2006(12).

[8]孙道功,李葆嘉.词汇—句法语义贯通研究的新探索[J].语言文字应用,2009(2).

[9]孙道功,李葆嘉.动核结构的“词汇语义—句法语义”衔接研究[J].语言文字应用,2009(2).

[10]周明海,亢世勇.语义角色句法实现的词汇语义制约信息库的建设及其应用[J].中国计算语言学研究前沿进展,2011(8).

[11]周明海.核心语义角色句法实现的词汇语义制约[D].烟台:鲁东大学,2011.

[12]周明海.词汇语义在语义角色句法实现的作用研究综述[J].语文学刊,2011(3-4).

[13]段弯弯.基于语义知识库的基本角色范畴句法实现的语义制约研究[D].南京:南京师范大学,2016.

[14]张晨.词汇语义制约语义角色映射为句法成分的特点及新词语语义推测研究[D].烟台:鲁东大学,2016.

[15]田震.非核心语义角色句法实现的词汇语义制约[D].烟台:鲁东大学,2014.

(下转第75页)

be divided into discrete and dimensional emotion databases, which have their own basic characteristics, advantages and disadvantages. The future direction of database improvement is proposed: increasing the background authenticity and emotional diversity of facial expression database, improving the manner of emotion induction and coding efficiency of micro expression database, and perfecting the evaluation quality of discrete emotion database.

Key words: discrete emotion database; dimensional emotion database; expression recognition; emotion induction

(责任编辑　合　壹)

(上接第 59 页)

Research on Language Feature Engineering for Chinese Automatic Analysis

FENG Zhiwei[1], CHENG Yong[2]

(1. School of Foreign Languages, Hangzhou Normal University, Hangzhou 311121, China;
2. School of Literature, Ludong University, Yantai 264039, China)

Abstract: The research on language feature engineering is the key to realization of Chinese automatic analysis, which includes "static features" describing inherent characteristics of words and the "dynamic features" describing word association in specific sentences. How to derive dynamic features from static features automatically is a difficult problem in automatic analysis of computer. Based on the large-scale annotation corpus, by systematically teasing out the complex linguistic feature sets in the process of Chinese automatic analysis, the paper summarizes the restriction rules about syntactic components, semantic roles and lexical semantics, which can guide the computer to realize the automatic derivation from static features to dynamic ones, so as to provide theoretical guidance and practical support for realizing Chinese automatic analysis.

Key words: static feature; dynamic feature; restriction rule of lexical semantics; Chinese automatic analysis

(责任编辑　梅　孜)