余小鹏, 徐健儿, 王振佩,等. 汉语框架语义网中词元语义搭配模式确定规则研究[J]. 智能计算机与应用,2024,14(8):29-31. DOI:10.20169/j.issn.2095-2163.240805

汉语框架语义网中词元语义搭配模式确定规则研究

余小鹏[1], 徐健儿[1], 王振佩[1], 姚小桐[2]
(1 武汉工程大学 管理学院, 武汉 430205; 2 武汉工程大学 计算机科学与工程学院, 武汉 430205)

摘 要: 机器理解自然语言和抽取信息涉及对词汇语义信息和句法信息两项重要内容的理解。汉语框架语义知识库(CFN)的词元语义搭配模式记录了词元与框架元素、词元与句法的组合方式,为机器自动进行语义分析和信息抽取提供了资源。而目前的词元语义搭配模式的数量多,机器确定并应用词元语义搭配模式存在一定困难,且框架元素间的关系比较模糊,词元配价成分的语义信息存在歧义,使机器不能准确理解和抽取文本信息。采用基于规则的方法,根据词元配价成分的语义信息和语义特征、词元外部的语境和框架元素提炼规则,可以使机器能够更准确地理解与抽取文本的语义信息。

关键词: 语义知识库; 框架语义; 规则; 词元语义搭配
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)08-0029-03

Research on determining rule of lexical semantic collocation patterns in Chinese FrameNet

YU Xiaopeng[1], XU Jianer[1], WANG Zhenpei[1], YAO Xiaotong[2]
(1 School of Management, Wuhan Institute of Technology, Wuhan 430205, China;
2 School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China)

Abstract: Machine understanding of natural language and information extraction involves the understanding of lexical semantic information and syntactic information. The lexical semantic collocation patterns of Chinese FrameNet record the combination of lexical units and frame elements, and lexical units and syntax, which provide resources for automatic semantic analysis and information extraction.There are a large number of lexical semantic collocation patterns now, which is difficult for the machine to determine and apply, and the relationship between frame elements is vague, and the semantic information of valency components of lemma is ambiguous, which makes the machine cannot accurately understand and extract text information. According to the semantic information and semantic features of the valence components of the lexical units and the external context of the lexical units and frame elements, the rules are extracted to make the machine understand and extract the semantic information of the text more accurately.

Key words: semantic database; frame semantics; rules; lexical semantic collocation

0 引 言

机器对语言理解涉及对词汇语义信息和句法信息两项重要内容的理解。汉语框架语义知识库(Chinese FrameNet, CFN)是在山西大学刘开瑛教授主持下,以框架为单位来描述词义、句子意义和文本意义,支持词一级语言单位的语义研究,能够表达特定情境的语义结构,可为机器提供词汇语义资源[1]。CFN 以框架语义学为理论基础,以 FrameNet 框架语义知识库提供的数据为参照而构建,由框架库、句子库和词元库 3 个部分组成[2]。CFN 框架库中的词元能够激活框架,并将框架所包含的框架元素与其词汇库的词元语义搭配模式相连接,词元与框架元素的组合方式及相应的词元与句法的组合方式,这种具体的“形式”特征可供机器使用,为机器自动进行语义分析和信息抽取提供了基础性资

基金项目: 教育部人文社会科学研究规划基金项目(19YJA880077)。

作者简介: 余小鹏(1974-),男,博士,教授,主要研究方向:信息系统与电子商务,数据挖掘,教育信息技术; 徐健儿(2000-),女,硕士研究生,主要研究方向:信息系统与电子商务,数据挖掘。Email:1334335720@qq.com; 王振佩(1999-),女,硕士研究生,主要研究方向:信息系统与电子商务,数据挖掘; 姚小桐(1999-),女,硕士研究生,主要研究方向:自然语言处理,数据挖掘。

收稿日期: 2023-05-19

源[3-4]。因此,词元的语义搭配模式对机器理解与抽取文本的语义信息尤为重要。

由于中文文本表述方式灵活多样,CFN 的训练数据规模有限,导致词元语义搭配模式不能满足应用的需求,使得机器应用词元语义搭配模式资源理解语言存在一定的困难。因此,本文引进、吸收 CFN 的研究成果,为了进一步满足中文信息处理应用的需求研究词元语义搭配模式。

1 提出问题

汉语框架语义知识库的词元语义搭配模式汇总报告记录了词元与框架元素的组合方式及相应的词元与句法的组合方式。其中,框架元素可使机器理解词汇的语义信息,但也不能保证机器能够正确理解语言,因为词汇需要按照一定的语义搭配组成句子,只有正确理解这些语义搭配,才能把不同的词汇的语义信息组成完整的语义概念[5]。

中文文本表述方式灵活多样,CFN 的训练数据规模有限,导致了词元语义搭配模式数量多,使机器确定并应用词元语义搭配模式存在一定困难,且词元匹配模式只记录了语义角色的短语类型、句法功能和框架元素,框架元素间的关系比较模糊,使得机器应用词元语义搭配模式资源理解语言和进行信息抽取时仍然存在一定的困难。明确框架关系有助于缓解标注语义角色时训练数据规模有限问题[6]。同时,由于框架元素类型的细化能够表示更丰富、更深化的语义信息,使得某些词元的配价成分的语义信息存在歧义[7],导致机器不能准确理解语言[7-8]。

以代数应用题领域词元的语义搭配模式为例,“这次考试成绩比上次考试成绩增加了 60%”,“这次考试成绩”和“上次考试成绩”被标注的框架元素都是属性(Att),但机器却难以理解这是哪个实体的属性,或者难以理解该属性是属于初值的属性还是终值的属性,同时,仅仅依靠“60%”的“短语类型”、“句法功能”和“框架元素”,机器就会默认“60%”与“这次考试成绩增加了 20 分”中的“20 分”具有相同的含义,但机器理解应用题题意后,需要对应用题进行求解,这些歧义现象会导致机器直接对“60%”进行加减运算。

从性质上可以将语言知识分为基于范畴的“属性:值”型知识和基于规则的“条件-->动作”型知识。其中,范畴表述的是对象的特征,规则表述的是范畴之间的关系[9]。从范畴和规则角度出发,细粒度剖析语言成分,可以增强搭配成分的语义知识的深度和广度,并且基于规则的方法采用可以解释歧义行为或歧义特性,使得机器能够更准确地理解与抽取文本的语义信息[9-10]。

综上分析,本文基于规则对词元语义搭配模式等进行延伸研究,根据词元构建规则,增强搭配成分的语义知识的深度和广度,使机器能更准确的抽取语义信息。本文在 CFN 的基础上,以代数应用题作为研究对象,阐述规则的构建方法,并可以将该构建方法应用到其他的领域。

2 词元规则的构建

2.1 规则的构建思路

根据词元的搭配对规则进行提炼的主要思路如下:

(1)根据框架中的词元,利用依存句法识别词元的配价成分;

(2)将配价成分进行统计和分类;

(3)筛选需要提炼规则的配价成分;

(4)根据配价成分的语义信息和语义特征、词元外部的语境和框架元素,即根据词元搭配模式提炼相应规则,构建规则库。

2.2 规则的形式化描述

根据配价成分的语义信息和语义特征、词元外部的语境和框架元素来提炼规则,需要更多、更复杂的逻辑关系表达,因此,本文采用 IF-THEN 语句来表示规则,形式如下:

IF <条件表达式> [<条件表达式>]…

THEN <Rule>

该语句的含义是,根据 IF 子句的条件表达式,在 THEN 子句中标注相应的规则;同时,一条规则中可以多次使用 IF-THEN 语句。为了使机器能够准确和有效的理解条件和规则,IF-THEN 语句中除了应用逻辑表达式符号,还采用自定义的符号,具体见表 1。

2.3 规则的示例及应用

本文以代数应用题为例介绍规则。代数应用题题目文本属于叙事类短文本,其文法、句法较为简单,情境特征非常明显,词语搭配的规律性比较强。针对代数应用题上述特征,从规则的角度出发,对词元的语义搭配规则进行提炼,为句法分析、语言成分之间的搭配提供更大的支持,使得机器能够更准确地抽取与理解文本语义信息。以下以“量变”框架中词元的部分规则为例,说明词元规则及其应用。

表 1　规则的常用符号

Table 1　Common notation for rules

符号	含义	
比较运算符	=	等于
	! =	不等于
多重条件	AND	和
	OR	或
	NOT	否
匹配	LIKE	如某种模式
	EXIST	存在
	IN	属于
	IS	是
其它符号	[tgt|word].LEFT	目标词或某个词语的左边
	[tgt|word].RIGHT	目标词或某个词语的右边
	[Frame].lexUnit	框架 Frame 中的词元

如果目标词(target, tgt)为“增加、增长、提高、上升、升、增”等表示属性值增加的词,能激活“量变”框架,tgt 后面的词语的语义特征为分数、百分数或者倍数,且被标注的框架元素为变幅(diff)时,此时机器不能确定多对属性与属性值之间的关系,也不可直接在目标实体的属性值上进行加减运算,而是需要根据词元的相应规则,确定属性与属性值之间的对应关系,并对目标实体的属性值进行更改,实现更准确地理解题意。

应用题语句中出现词语“比”时,可分为两种情况:①题目中只存在一个实体,则描述的是同一实体的某个属性在不同时间或空间状态下的对比;②词语“比”前后存在两个实体,则描述的是该两个实体间某个相同属性的对比。本文只对同一实体的某个属性在不同时间下的对比进行讨论。

如果目标词 tgt 所在的句子中出现由词语“比”引导的一个实体的一个属性,其属性值存在且被标注为初值(val_1),此时需要提炼规则如下:

IF tgt IN “增加”.lexUnit
　AND diff IS (fraction OR percent OR multiple)
　AND tgt.LEFT EXIST “比”
　AND val_1 IS NOT null
THEN val_2 = val_1 * (1+diff)

例如“某县去年植树造林 80 公顷,今年植树造林比去年植树造林增加了 25%”,根据提炼规则,机器可以获取到“去年植树造林”对应的属性值是初值(val_1)为 80,“今年植树造林”对应的属性值是终值(val_2),变幅(diff)是百分数,值为 25%,计算得出 val_2 的值为 100。

3　结束语

本文以框架语义理论为基础,以代数应用题领域词元的语义搭配为例,基于规则对词元语义搭配模式等进行延伸研究。由于目前的词元语义搭配模式的数量多,机器确定并应用词元语义搭配模式存在一定困难,框架元素间的关系比较模糊,并且词元的配价成分的语义信息存在歧义,使得机器应用词元语义搭配模式资源理解语言和进行信息抽取时仍然存在一定的困难。本文通过构建规则,增强搭配成分的语义知识的深度和广度,使机器能够更准确地理解与抽取应用题文本语义信息。此外,本文为其他领域增强搭配成分的语义知识的深度和广度提供了规则的构建方法,可以将该方法应用到其他的领域。

参考文献

[1] YOU L P, LIU K Y. Building Chinese Framenet Database[C]// Proceedings of the Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2005:301-306.

[2] LI R, LI S, ZHANG Z. The semantic computing model of sentence similarity based on Chinese FrameNet[C]//Proceedings of Web Intelligence/IAT Workshops. IEEE, 2009: 255-258.

[3] LI Ru, LIU Haijing, LI Shuanghong. Chinese frame identification using T-CRF model[C]// Proceedings of International Conference on Computational Linguistics .IEEE, 2010:674-682.

[4] 郝晓燕,刘伟,李茹,等. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报,2007,21(5):96-100,138.

[5] 徐晓东,刘昌. 句子理解的关键——对句法和语义关系的再探讨[J]. 心理科学进展,2008,16(4):532-540.

[6] 王晓晖,李茹,王智强,等. 基于 Self-Attention 的句法感知汉语框架语义角色标注[J]. 中文信息学报,2022,36(10):38-44.

[7] 由丽萍,白旭云. 汉语框架语义知识库非核心框架元素识别规则研究——以介词结构为例[J]. 情报学报,2011,30(12):1274-1279.

[8] 吕国英,武宇娟,李茹,等. 基于汉语框架语义的共指消解研究[J]. 计算机工程,2020,46(10):74-80,87.

[9] 詹卫东. 面向中文信息处理的现代汉语短语结构规则研究[D]. 北京: 北京大学,1999.

[10] 王鹏,戴新宇,陈家骏,等. 基于规则的汉语句法分析方法研究[J]. 计算机工程与应用,2003,39(29):63-66,169.