

基于语义角色和概念图的信息抽取模型

杨选选, 张 蕾

(西北大学 信息科学与技术学院, 西安 710127)

(y. xuanli@163. com)

摘 要: 传统的信息抽取方法由于缺少语义信息的支持, 抽取的准确率不高。针对这个问题提出了一种基于语义理解的信息抽取方法。一方面, 把语义角色标注的浅层语义信息转换成概念图, 无歧义地将抽取信息所包含的基本语义形式化; 另一方面, 通过概念图的相似度计算区分场景, 并使用语义角色获取抽取模式, 以提高抽取质量。实验结果表明, 该方法取得了较好的效果。

关键词: 信息抽取; 语义角色; 概念图相似度; 知网; 文本理解

中图分类号: TP311.5; TP391.2 **文献标志码:** A

Information extraction based on semantic role and concept graph

YANG Xuan-xuan, ZHANG Lei

(College of Information Science and Technology, Northwest University, Xi'an Shaanxi 710127, China)

Abstract: Because the traditional information extraction approaches are lack of semantic information, the accuracy is not high in extraction. In order to solve the problem, this article proposed a novel method of information extraction based on semantic role and concept graph. On one hand, the process transformed the shallow semantic information into concept graphs, and accurately described the main meaning of sentences. On the other hand, the calculator computed the similarity of concept graphs so as to distinguish the different domains of information. Meanwhile, the mapping rules would be got by using semantic role for improving the quality of extraction. The experimental results show that this method of information extraction is feasible and effective.

Key words: information extraction; semantic role; similarity of concept graphs; HowNet; text understanding

0 引言

信息抽取(Information Extracting)是以自然语言文档作为输入,产生固定格式、无歧义的输出数据^[1],它的一般功能是根据预先设定的任务,抽取特定类型或用户感兴趣的信息。处理的关键环节主要包括领域场景的识别、抽取规则的设计以及填充抽取模板。MUC(Message Understanding Conference)和NIST组织的自动内容抽取(Automatic Content Extraction, ACE)会议一直推动着这一领域的发展。有关中文的信息抽取研究起步较晚,并且由于汉语本身的特点,目前尚处于探索阶段。以往的研究工作主要集中在对同一领域信息的抽取,系统的扩展性和可移植性比较差,跨领域抽取大多数采用基于统计的方法,缺少对抽取文本的理解,抽取质量不高。而且由于汉语和印欧语系的差别,汉语词类跟句法成分之间不存在简单的对应关系,如果单纯地使用以语法分析为主的方法,势必会影响信息抽取的效果。近年来,随着研究工作的深入,在抽取中虽然也使用了一些语义信息,比如本体关系匹配的方法,但使用的语义主要还是以词法分析和独立的词义分析为主,并没有将信息中实际的语义引入进来。

知网^[2]是我国董振东先生研究创建,是以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念属性之间关系的知识库,并使用义原来标注概念。义原是知网中最基本的、不可再分割的意义最小单位。另外,它的

组建中还包括概念相关度计算器(Concept Relevant Calculator)和概念相似度计算器(Concept Similarity Measure),为中文的自然语言处理提供了良好的资源平台。

本文针对目前信息抽取存在的不足,提出把语义信息引入抽取过程,以语义角色和知网为基础,通过概念图将自然语言形式化,为抽取提供较为准确的语义信息和场景信息,从而提高抽取的准确性。

1 语义角色标注与概念图理论

在传统的自然语言处理过程(文本信息)中,一直以追求对文本语句的“深层理解”和“全面理解”为目标,但在纷繁的语言现象下,这种思想收效甚微。近年来,关于浅层语义分析^[3]的研究广泛展开,并取得了较多的研究成果,语义角色标注^[4]就是其中之一。概念图^[5]是一种知识表示工具,语义角色反映了句子的基本语义信息,通过概念图可以比较准确地将语义角色标注的基本语义形式化、可计算化,为后面的语义计算和信息抽取提供基础。

1.1 语义角色标注

语义角色(Semantic Role)是指有关语言成分的所在语句所表达的事件中所扮演的参与者角色(Participant Role),是语言学家对句子中有关结构成分之间的意义关系的一种分类,它在一定程度上反映句子的意义,并将其形式化。语义分析是自然语言理解领域的根本性问题,限于目前的技术和理论

收稿日期:2009-08-18;修回日期:2009-10-12。

作者简介: 杨选选(1984-),男,陕西周至人,硕士研究生,主要研究方向:人工智能、自然语言理解; 张蕾(1964-),女,陕西西安人,教授,博士,主要研究方向:人工智能、自然语言理解。

水平,深层的语义分析很难做到,但在某些应用场合,浅层语义分析(Shallow Semantic Parsing)——一种简化了的语义分析方式——确实能够取得比较好的应用效果。它不考虑时态变化和枝节性的信息,只专注于谓词所支配的基本句子意思。

在实际语句中,语义角色指示句中谓词与它的参数之间的语义关系。语义角色标注(Semantic Role Labeling)就是将词语序列分组,并按照语义角色对它们进行分类,它是浅层语义分析的一种重要实现方式,它体现了句子的基本意思。该方法并不对整个语句做详细的语义分析,而只是标注句子中给定谓词(动词、名词等)的语义角色(参数),从而使计算机对语句有一个“浅层”的理解。比如:[运动员(Agent)][一周后(Tmp)]将要[参加(V)]正式[比赛(Patient)]。这句话中,“参加”是谓词,“运动员”和“比赛”分别是其“施事者”和“受事者”,“一周后”表示其发生的时间。此句也可表示为:[一周后(Tmp)][运动员(Agent)]将要[参加(V)]正式[比赛(Patient)]。这两句谓词“参加”在各句中所扮演的语义角色相同,也就是说虽然它们的表述形式不一样,但它们的意思相同。

这是一个简单的例子,但是它告诉我们谓词及其参数在句子中表达语义的重要性。我们可以通过语义角色中标明的时间、地点、施事、受事等角色和领域信息内容的特点,进行针对这些类信息的提取。

1.2 语义角色参数

谓词的语义角色分为核心语义角色和附加语义角色^[6],核心的语义角色一般表示为 ARG 后直接跟数字,表示的意义根据谓词的不同而有所差异,幸运的是对于汉语中的大多数谓词来讲,其角色参数在句中比较固定,ARG0 一般表示其施事者,ARG1 表示动作结果的受事者。这种形式为后面的概念图生成提供了便利条件。其余的语义角色为附加语义角色,用前缀 ARGM 表示,后面跟一些附加标记来表示这些参数的语义类别,如 ARGM-LOC 表示地点,ARGM-TMP 表示时间等。

1.3 概念图理论

概念图是一种有力的知识表示工具,能完全描述自然语言所表达的意思,实现与自然语言的互译。与传统的知识表示方法相比较,概念图更具表达力与可读性,且能与自然语言相互映射,它有两个节点,一个是概念节点一个是关系节点。概念节点表示一个具体或抽象的概念;关系节点则表示概念之间的关系,比如处所(LOC),时间(PTIM),相关(INST)等。

概念图具有严密的数学结构,可以进行拷贝、限制、连接和化简操作。正是由于这些特点,使得概念图在表示自然语言时能有效地避免歧义,准确地反映句子的意思。另外,可以通过概念图的匹配来衡量它们所表示的两个句子意义间的相似程度,传统的匹配方法包括投影匹配和最大连接匹配,它们同属于不完全匹配,如果匹配失败,则无法衡量两个句子的相似程度。本文通过概念图的相似度^[7]计算来完成概念图的匹配。

$$\left\{ \begin{aligned} SoG(c_Q, c_R) &= w(c_Q, c) \times sim_e(c_Q, c_R) + \\ &\max \left\{ \sum_j w(c_Q, j) \times sim_r(r_Q^j, r_R^j) \times [SoG(c_R^j, c_R^k)] \right\} \\ w(c_Q, c) + \sum_j w(c_Q, j) &= 1 \end{aligned} \right\} \quad (1)$$

$$sim_r(r_Q, r_R) = \begin{cases} 1, & r_Q \supseteq r_R \\ 0, & \text{其他} \end{cases} \quad (2)$$

其中: sim_r 表示概念图关系的相似度, sim_e 表示概念的相似度,本文使用知网的概念关联计算模块计算其值; c_Q 和 c_R 分别表示查询图和资源图的入口点; r_Q^j 和 r_R^j 分别表示查询图和资源图中的第 j 条关系; c_Q^j 和 c_R^k 分别对应其入口点; $w(c_Q, c)$ 表示入口权值, $w(c_Q, j)$ 表示和入口相关的第 j 条关系的权值。利用式(1)可以计算概念图的相似度。

2 基于语义角色和概念图的信息抽取

在跨领域的信息抽取系统中如何将待抽取信息进行准确的领域场景划分以及获取相关领域的抽取规则是整个抽取方法需要解决的问题。在本文提出的抽取方法中,对待抽取的信息进行相关的预处理和语义角色标注,将语义角色标注的语义信息转换为概念图,通过对抽取信息的语义计算来识别领域场景,并根据语义角色获取抽取规则。最后通过与规则库中相关领域的抽取规则进行匹配,映射信息点完成抽取。

2.1 领域场景划分的解决方案

在信息抽取中,不同的领域场景往往对应不同的抽取模板和抽取规则,是否能够正确地识别场景,对后续的抽取工作有着重要意义,它是跨领域抽取不可回避的问题。本文提出的思想是:根据语义角色构建概念图并结合知网信息,通过与领域概念图图库中的领域概念图进行相似度计算,进行场景的划分。生成概念图的做法是:对待抽取信息进行语义角色标注,将语义角色成分映射为概念图中的概念节点,语义角色是句中语义成分与句中谓词关系的表示,通过提取这种关系,并按照一定的转换规则,将语义角色转换为概念图中的关系并体现在概念图中。这里使用的谓词角色参数均采用 Chinese PropBank 中对语义角色的定义,表 1 为概念图关系和语义角色的转换关系(“—”表示不转换)。

将句中的语义信息转换为概念图后,对这种含有语义信息的信息的概念图进行相似度计算,从而实现领域场景的划分。

考虑句 1 “[ARG0 北京][ARGM-TMP 明天][TARGET 有][ARG1 小雨]。”将语义角色与概念图按照对应关系进行转换,如图 1 所示。

表 1 概念图语义转换关系

语义角色	概念图关系	语义角色	概念图关系
ARG0	AGENT	ARGM-MOD	—
ARG1	RCPT	ARGM-TMP	PTIM
ARGM-DIR	—	ARGM-DIS	INST
ARGM-LOC	LOC	ARGM-PNC	—
ARGM-REC	AGENT; RCPT	ARGM-EXT	MANR
ARGM-CAU	CAUS	TARGET	入口点
ARGM-ADV	MANR	ARGM-NEG	—

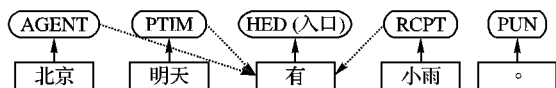


图 1 句 1 转换后的概念图

由于自然语言的灵活性,同一领域的信息可以有多种表达方式,尽管它们在形式上不同,但它们在基本的语义层面上是相同或相近的。为了抽象出相同领域信息的这种语义相似性,我们采用知网作为语义标注的词汇库,在构建具有一般意义的领域概

念图时,结合知网,将标注的词语与知网中的首义元进行替换。知网中“阴天”定义为:DEF = { Weather | 天气; CoEvent = { WeatherBad | 坏天气 } },“晴天”定义为:DEF = { Weather | 天气 }。“晴天”和“阴天”它们表达的意思都是“天气”。

通过首义原替换,进一步将图1表示成图2的具有一般意义的领域概念图形式,可以将其看作成一个简单的天气领域概念图。

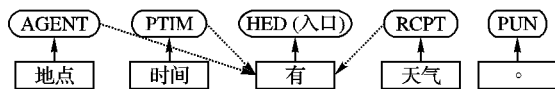


图2 句1替换后的义原概念图

句2“[ARGM - TMP 6月28日], [ARG0 重庆] [ARGM - ADV 阴天] [TARGET 转] [ARG1 多云]。”的概念如图3所示。通过将表示待抽取信息语义的概念图和领域概念图进行相似度计算确定抽取句子的领域场景,把计算的概念图相似度做降序排列,按照相似度的大小划分领域场景,如果计算的相似度值均小于0.5,则认为场景匹配失败。

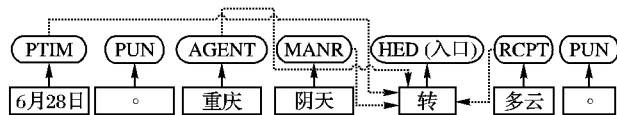


图3 句2转换后的概念图

概念节点的相似度结合知网的 Concept Similarity Measure 模块计算。计算图3与图2两个概念图的相似度 $SoG(c_3, c_2)$ 。计算如下:

$$sim_1 = 0.4 \times sim_c(\text{重庆, 地点}) + 0.6 \times sim_r(\text{AGENT, AGENT}) = 0.4 \times 1 + 0.6 \times 1 = 1$$

$$sim_2 = 0.4 \times sim_c(\text{重庆, 时间}) + 0.6 \times sim_r(\text{AGENT, PTIM}) = 0.4 \times 0.4 + 0.6 \times 0 = 0.16$$

$$\max(sim_1, sim_2) = 1$$

同样地按相同方法计算后续的概念关系,按照式(1)计算得出 $SoG(c_3, c_2) = 0.92$ 。与其他概念图相似度计算类同,如果计算结果都小于0.92,那么将句2划入天气领域。

2.2 抽取模板

世界上的事物各自具有不同的属性,但不同属性之间往

往具有一定的规律性和联系。这种规律性的知识经过提炼,就可以形成人们认识某一类事物的一种固定的框架^[8]。当描述一个事物时,从已知的框架库中找出一个合适的框架,根据实际情况对其细节加以修改、补充,从而形成对当前事物的完整描述。

一般情况下,相同场景的信息在描述方式以及内容方面有着较强的相似性,比如:天气场景的信息包括某一时间,某一地点的天气状况,以及温度湿度等方面的信息。根据这个特点,用适用于文本信息的层级框架模板对其进行描述。在同一领域中,框架包含的侧面名基本相同,生成框架时,为每一个信息槽构建一个指示器(Point),以便于指向抽取文本的信息点和语义角色,如图4所示。

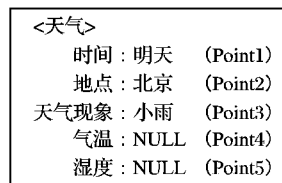


图4 句1的框架描述

图4中,Point1, Point2, Point3 分别指向句中的 ARGM - TEPMP, ARG0, ARG1。因为句中没有气温和湿度的信息,所以Point4 和 Point5 为空。

2.3 获取映射关系

图4给我们展示了一个有趣的现象:句子中的语义角色成分,以某种对应方式和抽取模板中的信息槽之间存在对应关系,而这种语义角色与框架槽之间的映射关系在不同领域是不相同的。表2表示了不同领域的不同映射规则。这种映射关系可以看作是场景要素与现实文本信息点的抽象,场景模板中的不同槽分别代表不同的语义角色,语义角色可以表示场景要素。

在获取映射规则时,我们构建的训练数据包括两部分:1)领域相关的文本信息;2)与文本相对应的框架模板。每一个框架模板中的指示器指向抽取的资源文本,其形式与图4类似。当一个完整的语义角色结构被区分出来时,根据框架的侧面名和槽建立映射关系,并收集规则。

表2 映射关系

场景	映射规则
天气 场景	1) 地名 & ARG0 → 地名槽(ARG0)
	2) ARGM - TMP + TARGET{ 转 有 出现 到 天气变化方式 } → 时间槽(ARGM - TMP)
	3) TARGET{ 转 有 出现 到 天气变化方式 } + ARG1 → 天气状况槽(ARG1)
	4) ARGM - ADV + NN + 数词 → 气温槽(数词)
招聘 场景	1) 名词 & ARG0 → 聘用单位槽(ARG0)
	2) 名词 & ARG1 → 职位槽(ARG0)
	3) (ARGM - LOC ARGM - TMP) & 地点 → 工作地点槽
	4) ARGM - ADV + 数词 → 薪水槽(数词)

3 算法

基于语义角色与概念图的信息抽取流程如图5所示。

算法步骤如下:

1) 对于处理的文本,根据香港科技大学提供的 Chinese Semantic Parser 语义角色标注系统进行语义角色标注和词性标注;

2) 将1)处理的结果根据本文中提出的方法,把语义角色

转化为概念图中的语义关系,建立相应的概念图;

3) 通过本文中说明的概念图相似度计算办法,进行领域场景的识别,如果识别成功进入下一步处理,否则进入待处理领域,等待处理;

4) 进行相关领域的映射规则匹配,若匹配成功进入5)处理,否则算法结束;

5) 根据4)中的匹配结果填充抽取模板,将信息点映射到模板槽中,生成抽取结果。

4 实验结果与分析

本文对“天气场景”和“招聘场景”的信息进行实验抽取,实验数据选自1998年人民日报语料库的部分数据和实验室收集的数据共166篇相关领域的文章和信息,共计1253条例句。实验中,随机选取400条作为测试数据,其余853条信息作为训练数据。系统评价采用准确率(*precision*)、召回率(*recall*)和*F-measure*对模型的有效性进行分析。准确率等于系统正确抽取的结果占有所有抽取结果的比例,召回率等于系统正确抽取的结果占有所有可能正确结果的比例,*F-measure*计算如式(3),实验结果如表3所示。

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

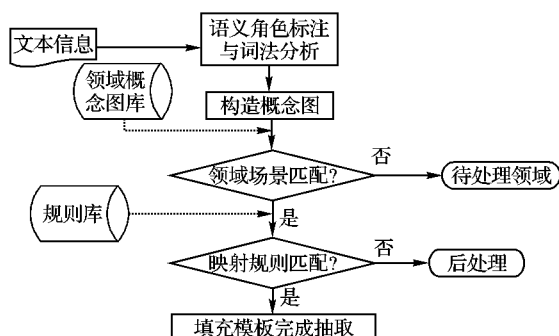


图5 基于语义角色和概念图的抽取模型

表3 实验结果

抽取技术	Precision/%	Recall/%	F-measure/%
基于语义角色和概念图	78.59	83.12	80.79
基于本体关系匹配	76.93	81.16	79.00
基于统计	75.25	81.06	78.05

实验显示采用基于语义角色和概念图的抽取技术,提高了信息抽取的准确率和召回率,这主要是本文的方法在场景识别和获取抽取模式时获得了一定的效果。

通过实验分析,语义角色本身代表了一个语义块,与词性和其他一些语法单位相比较,它是一个真正含有语义的信息单位,通过转换,概念图可以清晰地表示语义中的概念和它们

之间的关系,在场景识别中,使用这种方法区分场景得到的准确率要高于传统的统计识别,增强了系统的可扩展性。另外,通过使用语义角色获取抽取规则,不仅在一定程度上克服了由于语义信息的缺失造成的抽取准确率不高,而且还有效地避免了多语言单位混合使用引发的语言单位间的互相干扰,从而改善了抽取质量。

5 结语

本文提出的基于语义角色和概念图的信息抽取模型,是在语义层面对信息抽取的尝试,它将浅层的语义信息应用于场景识别和抽取模式两个层次上,并通过概念图将句子的语义形式化、可计算化。概念图义原与句子语言单位间的替换,使得领域概念图更具一般性,提高了场景识别的准确率,增强了系统的可扩展性。从实验结果可以看出本文所提出的抽取模型和抽取策略是可行且有效的。

参考文献:

- [1] 刘迁,焦慧,贾惠波. 信息抽取技术的发展现状及构建方法的研究[J]. 计算机应用研究, 2007, 24(7): 6-8.
- [2] 董振东,董强. 知网的理论发现[J]. 中文信息学报, 2007, 7(4): 36-43.
- [3] 陈耀东,王挺,陈火旺. 浅层语义分析研究[J]. 计算机研究与发展, 2008, 45(S1): 321-325.
- [4] 袁毓林. 语义角色的精细等级及其在信息处理中的应用[J]. 中文信息学报, 2007, 21(4): 10-11.
- [5] 张蕾,李学良. 概念结构及其应用[D]. 西安: 西北工业大学, 2001.
- [6] XUE NIANWEN. Labeling Chinese predicates with semantic role[J]. Computational Linguistics, 2007, 34(2): 226-231.
- [7] ZHONG JIWEI, ZHU HAIPING, LI JIANMIN, et al. Conceptual graph matching for semantic search[EB/OL]. [2009-08-01]. <http://apex.sjtu.edu.cn/docs/iccs2002.pdf>.
- [8] 尹朝庆,尹浩. 人工智能与专家系统[M]. 北京: 中国水利水电出版社, 2002.
- [9] 周顺先. 文本信息抽取模型及算法研究[D]. 长沙: 湖南大学, 2007.
- [10] CALIFF M E, MOONEY R J. Bottom-up relational learning of pattern matching rules for information extraction[J]. Machine Learning Research, 2003, 4: 177-210.
- [7] ZHANG DENG-SHENG, LU GUO-JUN. Review of shape representation and description techniques[J]. Pattern Recognition, 2004, 37(1): 1-19.
- [8] CHEN BO, PAN XIANG. Geodesic Fourier descriptor for 2D shape matching[C]// Proceedings of the 2008 International Conference on Embedded Software and Systems Symposia. Washington, DC: IEEE Computer Society, 2008: 447-452.
- [9] ZHANG DENG-SHENG, LU GUO-JUN. Study and evaluation of different Fourier methods for image retrieval[J]. Image and Vision Computing, 2005, 23(1): 33-39.
- [10] MEI YE, ANDROUTSOS D. Affine invariant shape descriptors: The ICA-Fourier descriptor and the PCA-Fourier descriptor[C]// ICPR 2008: Proceedings of the 19th International Conference on Pattern Recognition. Washington, DC: IEEE Computer Society, 2008: 1-4.
- [11] EL-GHAZAL A, BASIR O, BELKASIM S. A novel curvature-based shape Fourier descriptor[C]// Proceedings of the 15th IEEE International Conference on Image Processing. Washington, DC: IEEE Computer Society, 2008: 953-956.
- [12] ZHANG DENG-SHENG, LU GUO-JUN. Shape-based image retrieval using generic Fourier descriptor[J]. Signal Processing: Image Communication, 2002, 17(10): 825-848.
- [13] JIA HAI-TAO, XIE MEI. Improvement of Fourier descriptor using spatial normalization[C]// ISCIT 2005: IEEE International Symposium on Communications and Information Technology. Washington, DC: IEEE Computer Society, 2005: 1284-1287.
- [14] LATECKI L J, LAKAMPER R, ECKHARDT T. Shape descriptors for non-rigid shapes with a single closed contour[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2000: 424-429.
- [15] JEANNIN S, BOBER M. Description of core experiments for MPEG-7 motion/shape, MPEG-7[S]. 1999.

(上接第363页)