基于记忆学习方法在自然语言处理中的应用

Memory-Based Learning for Natural Language Processing

鲁 松 白 硕

(中国科学院计算技术研究所 北京100080)

Abstract Natural Language Processing is in a new phase with introducing Statistical method and Machine Learning. This paper roundly introduces, analyzes and evaluates Memory-Based Learning(MBL), one method of Machine Learning, about representation, acquirement and reasoning of Natural Language knowledge, and points out the advantages and disadvantages of MBL and analyzes deeply their reasons. At last, a disambiguation experiment of Chinese Numeral by MBL is given.

Keywords Natural language processing, Memory-based learning, Analogical reasoning, Similarity computation

1 背景介绍

在诸多自然语言处理应用问题中,传统手工规则的失败暴露了经典人工智能IF-THEN推理模式在自然语言知识体系尚未完全把握的情况下的缺陷。由此,在一定知识推理机制体系下,自然语言的知识表示和知识获取作为自然语言处理中的关键问题已经成为困扰其形式化的主要瓶颈。

"规则+例外"的框架下,"重规则,轻例外"的处理方式在自然语言知识框架不清楚和难以确定的情况下并不适用于自然语言处理。更明确地讲,基于现有自然语言知识框架下的规则体系过于粗糙,难以适用于存在大量"例外"的自然语言现象,知识颗粒度不细,泛化过强,是造成自然语言处理中大量歧义出现的主要原因。

基于大规模语料训练的统计方法在语音识别领域[1,16,17]取得巨大成功后,被引入自然语言处理之英文词性标注(Part of Speech)[6]问题中,同样取得了大幅度的改进,这一成功激励了统计方法在自然语言处理其他问题的迅速推广。

我们认为统计方法的成功之处在于它是一种以"数量"的方式来表示解决特定自然语言问题中所需相关个别事例间的自然语言知识,尽管这种以"数量"形式的知识表示可理解性不强,但相对于现有各种概念类(class)为单位的自然语言知识而言却是"细腻"的。但与此同时,这种基于相关个别事例的知识表示方法由于缺乏概念层次上的泛化能力,也不可避免地遇到覆盖能力不足的问题,即:数据稀疏问题。尽管统计方法在多种应用问题中取得了很大的成绩,但其知识表示的不可理解性、数据稀疏问题和非增量式的缺点在一定程度上限制了它的应用。

与此同时,各种经典的机器学习方法开始被引入自然语言处理中,采用决策树ID3方法[Masahiko, 1999]和归纳逻辑程序方法ILP(Inductive Logical Programming)[15]构建句法分析器,Exemplar-based learning的词义消歧[4],以及应用于词性标注[4],组块分析(Chunk Parsing),语调标注,英语中的介词短语问题(Prepositional Phrase)[14]的基于记忆的学习方法(Memory-Based Learning)[5]等。

这些机器学习方法所依托的知识表示体系主要有两种,一是基于属性逻辑(attribute-logic)的表示方法,如ID3方法和Exemplar-Based方法;二是基于一阶谓词逻辑的表示方法,如ILP方法。知识表示体系与知识推理机制密切相关,不同的知识表示体系决定了不同的推理机制,但相同的知识表示体系同样可以使用不同的推理机制。

其中基于属性逻辑表示体系下相似度类比推理的MBL学习方法的突出特点是在知识获取过程中不做任何形式概念提升的泛化工作,知识获取的过程就是记忆数据的过程,针对自然语言处理中知识难以获取的困难,它的这一特点从机制上提供了一种有效的解决途径。正是由于这一点,本文将对MBL方法从知识表示、知识获取、知识推理以及其对自然语言处理的适用性等几个方面对MBL方法进行全面的描述、分析和评价,并通过MBL方法在汉语数词语义类消歧中的应用予以说明。

2 基于记忆的学习方法

MBL 方法是由基于记忆的推理(Memory-Based Reasoning,MBR)[5]演变而来的机器学习方法。这一推理模式确认的推理假设为,知识推理的过程是基于经验的相似度比较的过程,而不是基于归纳概念的条件-动作过程。在这一假设的基础上,决定了 MBL 方法的推理机制是相似度计算的类比推理,也决定了知识获取的过程不是概念的提升和归纳,而是以属性逻辑表示经验的记忆过程,这一特点使学习的过程和复杂度得到了大大的简化。

2.1 MBL 基本框架

MBL 方法的知识处理框架如表1所示。

表1 MBL 方法的知识处理框架

过程	方式
知识表示	属性逻辑(属性-属性值)
知识获取	记忆存储
知识推理	相似度比较

MBL 方法的推理方法源于模式识别中经典分类方法 k-NN 方法,其基本框架是:在待处理数据的处理过程中,将计算待处理数据与所有存储训练数据之间的距离,找到与待处理数据距离最近的 k 个训练数据,通过对 k 个结果的判决,为待处理数据指定距离最近样本点的所在类别,完成分类工作。

2.2 相似性比较机制

相似性比较机制是 MBL 方法中推理过程的基本操作,通过向量之间的距离计算来完成。其基本结构为:$X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_n)$ 为 n 个属性的两个向量,二向量的相似度计算通过距离计算来完成,形式如下:

$$\Delta(X,Y)=\sum_{i=1}^{n}w_i\delta(x_i,y_i) \tag{1}$$

其中 $\Delta(X,Y)$ 为向量 X 和向量 Y 之间的距离求解;w_i 是第 i 个属性在分类中的权重,反映属性 i 在分类中的贡献量;$\delta(x_i,y_i)$ 表示向量 X 第 i 个属性的属性值与向量 Y 第 i 个属性的属性值之间距离的计算求解。

由于 MBL 方法推理模式是基于距离计算进行的,因此推理结果对距离计算方法极为敏感[12]。为了获取精确的距离计算,在公式(1)的框架下,针对离散的自然语言符号,可以通过调整各属性权重 w_i 和属性值间距离计算方法来加以改进。其中下面的 IB1 和 IB1-IG 采用的是调整属性权重的方法,而 MVDM 是一种定义属性值间距离计算的方法。

IB1 方法 针对自然语言的离散符号,IB1 和 IB1-IG 两种方法的属性值距离计算 $\delta(x_i,y_i)$ 均被定义为式(2)形式:

$$\delta(x_i,y_i)=\begin{cases}0 & \text{if } x_i=y_i\\ 1 & \text{else}\end{cases} \tag{2}$$

且在 IB1 相似性比较方法中各属性权重 $w_i=1(i\in\{1,\cdots,n\})$,即各个属性对整个分类任务所起的作用是相同的。尽管此种距离计算方法简单,但针对一些分类问题,效果仍很理想。

IB1-IG 方法 在缺乏必要信息的情况下,IB1 中设 $w_i=1$ 的简单距离计算模型是恰当的,但利用统计信息和引入信息论中信息增益的概念可以基于经验确定各个属性的权重。由此,可自动获取各属性在分类任务中所起作用的大小,即各属性权重 w_i,此方法被称为 IB1-IG,如式(3)所示。

$$w_i=\frac{H(C)-\sum_{v\in V_i}P(v)\times H(C|v)}{s_i(i)} \tag{3}$$

式(3)的解释为:通过训练数据集中类别分布比例上的不确定性(信息熵)和在向量中已知某一属性的情况下类别不确定性(条件熵)减少量的计算结果来定义此属性的权重。

式(3)中的 C 是类别分布比例集合,V_f 是属性 f 中所有属性值的集合,其中的 $H(C)$ 是类别分布比例的不确定性,被定义为式(4):

$$H(C)=-\sum_{c\in C}P(c)\log_2P(c) \tag{4}$$

式(3)中的分母被定义为式(5)的形式,即标准化因子,用来避免权重值向属性值过多的属性倾斜。

$$s_i(i)=-\sum_{v\in V_i}P(v)\log_2P(v) \tag{5}$$

无标准化因子的方法被称为 InfoGain,添加标准化因子的方法为 GainRation。IB1-IG 确定向量中属性权重的技术不仅使基于属性逻辑的表示方法具有更强的表现能力和预测能力,而且在一定程度上具有针对数据稀疏问题的平滑能力[13]。

MVDM 方法 在 IB1 和 IB1-IG 方法中,属性值的距离计算仅是离散符号相同与否的判断,因此无法充分估计同义词或在特定自然语言问题中异形相关词语相似性的情况,为了解决这一问题,文[9]提出并由 Cost[1993]改进的 MVDM(Modified Value Difference Metric)属性值距离计算方法。此方法是针对特定问题及其相应训练集,通过基于属性值的类分布情况构造两两属性值间分布距离矩阵来实现属性值间的距离计算,计算公式如下:

$$\delta(V_1,V_2)=\sum_{i=1}^{n}|P(C_i|V_1)-P(C_i|V_2)| \tag{6}$$

式(6)中属性值 V_1 和 V_2 的距离为二属性值在各个类中分布概率差的累加值。

MVDM 方法可以说是在基于相似性推理的基础上,采用统计方法构造(或称学习)了一个针对特定分类问题、属性值概率相关性的知识库。尽管 MVDM 方

法在一定程度上实现了属性值之间的相关性概念，但其获取知识的特定任务相关性，同所有统计方法一样，不可避免的数据稀疏等问题限制了它的应用。

3 自然语言处理与基于记忆的学习方法

MBL 方法是一种针对分类问题的有导师学习方法，而自然语言处理中的许多问题都可以被形式化为典型的分类问题，例如词性标注、多义词消歧、语音合成中语调调整、NP 组块分析和英文中的 PP 问题（Preposition Phrase）等。针对自然语言处理，MBL 方法的优势和潜力主要体现在以下几点：

（1）无损的知识获取，概念的泛化和抽象往往会引进知识的不一致性，特别是在信息庞杂的自然语言中更是如此，而 MBL 方法不存在概念的泛化和抽象，在保证信息量不减少的同时，也就避免了知识提升中带来的不一致性。

相对于决策树 ID3及其改进版本 C5.0的剪枝过程和统计方法中小概率事件参数难以估计的缺陷，MBL 方法对训练实例仅进行记忆不做任何归纳的知识获取方法是保证知识无损获取的关键，由于自然语言中存在着大量的例外和难以明确形式化的知识与子类知识，给任何形式的抽象和归纳带来了障碍，因此训练数据以属性和属性值形式不做任何改变的存储这一特点在一定程度上缓解了自然语言知识表示困难和获取困难的问题，克服了"重规则，轻例外"给表示自然语言灵活性带来的限制。

（2）基于相似性推理的积极作用，自然语言中存在大量的非精确知识和模糊概念，这些知识的表示和获取是困难的。也正是因为统计方法在一定程度上解决了这一问题，所以它取得了巨大的成功。这一成功不仅来源于针对个别对象"细腻"的知识表示，而且来源于知识推理中量化的度量方法（Bayes'方法及 HMM 等）。这种基于实数域量化的程度度量方法，相对于"非此即彼"的一元逻辑来讲，对知识的推理则是更为精确的。而 MBL 方法中的相似度计算在这一方面具有与统计方法等同的作用，不仅如此，MBL 方法的相似度计算机制和属性权重计算的引入避免了统计方法中平滑数据稀疏中参数估计的诸多困难。

以上两点是 MBL 方法在自然语言处理的许多问题中取得了突出成绩的关键，其应用效果可以参见文[5]。

虽然 MBL 方法存在许多比其他方法更适合自然语言处理的特点，但问题和缺点仍然存在。

1）属性逻辑表示方法上的限制　基于属性逻辑的知识表示方法在医疗领域病情描述、制造业器件属性、天气预报的气象指标等以特定属性描述集合为知识的应用领域是恰当的，但对以语序灵活为主要特点的自然语言来讲，便会由于知识表示机制的限制带来数据稀疏问题，当以上下文的临近特定位置词语作为属性时更是如此[5]。

2）递归知识的不可学习　句法成分的递归性质是自然语言的基本特性，也是 Chomsky 短语生成语法的精髓之一，但 MBL 方法的属性逻辑知识表示方法从根本上限制了递归性知识的获取。不引入具有更强表示能力的一阶谓词逻辑表示方法，递归机制的学习就无法实现。

3）知识的冗余存在　自然语言中存在大量无法纳入"规则"的"例外"，但并不等于没有规律存在。而 MBL 方法在将所有知识视为"例外"进行无损知识获取的同时，以牺牲泛化为前提，放弃了对规律的提取和归纳，不可避免地造成自然语言知识库中冗余现象的存在，即：知识颗粒度过细的问题。由此带来的一个副作用就是推理效率低的问题。在 MBL 方法的框架下，知识颗粒度的大小与推理效率的高低成为一对矛盾。针对这一问题，文[5]提出了 IG-TREE 的方法构造近似最优树来提高搜索速度。

需要进一步指出的是，这些问题并不是 MBL 方法所独有的，在其他机器学习方法中，特别是在相同知识表示框架下，同样存在这些问题。

4 MBL 方法在数词词义消歧中的应用

在此将给出 MBL 方法的数词词义消歧中的一个应用实例。

汉语数词作为自然语言的重要组成部分，不仅占到了高达3.97%的比例，而且其描述功能已经不仅仅局限于简单的计数作用，是关系到语音合成、人机对话和信息提取等直接效果的重要因素。在机器翻译的研究中，数词消歧更是一个词语一级处理的主要任务之一[4]。

本文将该问题形式化为数词"一"和数词非"一"两个部分的分类问题，分别进行考虑，数词"一"具有9种常用义项，数词非"一"具有6种句法或语义功能。在有导师分类学习中，每个义项作为一个类别。属性向量被定义为数词上下文环境中词语及其词性：〈$word_{-1}$，POS_{-1}，numeral，$Word_{+1}$，POS_{+1}，$word_{+2}$，POS_{+2}，target sense〉

例如：〈民国，N，二十六，M，年，Q，春天，N，ordinal numeral〉，其中针对各自训练集计算相应权重属性，结果如表2所示。

实验中逐步增加训练集大小，以观察训练数据量大小对分类效果的影响，实验比较结果如表3所示。

表2 属性权重计算表

属性	数词"一"		其他	
	基于13,515个实例		基于11,618个实例	
	InfoGain	GainRation	InfoGain	GainRation
$word_{-1}$	0.377039	0.0776437	0.363583	0.0479474
POS_{-1}	0.236486	0.0753985	0.0961027	0.0292028
数词	Ignore	Ignore	Ignore	Ignore
$word_{+1}$	0.657096	0.10687	0.384262	0.055386
POS_{+1}	0.284081	0.221041	0.0857839	0.0388101
$word_{+2}$	0.501508	0.0483735	0.374327	0.0413261
POS_{+2}	0.0988993	0.0335989	0.0459102	0.0135676

表3 实验结果

	"一"的分类				数词非"一"的分类			
训练集	3,000	6,000	9,000	13,515	3,000	6,000	9,000	11,618
测试集	399	598	923	1,701	340	570	950	1,100
IB1-IG	46.93%	59.23%	73.01%	89.48%	56.98%	67.23%	83.75%	95.48%

实验结果反映出正确率随训练实例数的逐步增加而提高，最终结果较原手工规则分别提高了13%和14%，改进效果明显。

（限于篇幅，数词分类方法、举例及IB1-IG的属性权重计算结果等详细描述在此省略，可参见文[1]）

讨论与结论 本文就MBL方法在自然语言处理中的优缺点及其产生原因进行了详细的论述，并就MBL方法的知识表示和知识获取的一些特点与其他方法进行了针对性比较。

MBL方法的类比推理机制给针对自然语言知识的知识获取带来了极大的便利，在一定程度上克服了自然语言知识表示困难和知识获取困难的瓶颈，但仍然存在其他问题，"重规则，轻例外"不适于自然语言的灵活性，而完全抛弃规则，放弃规律性事物的归纳同样是偏激的，如何根据自然语言特定的性质在二者之间获取一个恰如其分的均衡将是至关重要的，这也是我们现在的工作。

参 考 文 献

1 鲁松，孙红梅，白硕 自然语言处理中记忆学习方法的改进.第六届计算机科学与技术研究生学术研讨会，2000

2 袁江，宋柔.汉语文语转换中的阿拉伯数字串读法初探.中文信息处理国际会议，北京，1998.136～142

3 Aha D W. Lazy learning·Special issue editorial. Artificial Intelligence Review，1997，11：7～10

4 Daelemans W，Zavrel J，Berck P，Gillis S. MBT：A memory-based part of speech tagger generator. In：E. Ejerhed，I. Dagan，eds. Proc. Of Fourth Workshop on Very Large Corpora，14-27，ACL SIGDAT，1996. 14～27

5 Daelemans W，et al. Forgetting Exceptions is Harmful in Language Learning. Machine Learning，Special Issue on Natural Language Learning，1998，11：1～3

6 Cutting D，Kupiec J，Pedersen J，Sibun P A practical part-of-speech tagger. In：3rd conf. on applied natural linguistic processing(ANLP-92)，1992 133～140

7 Rabinec L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE，1989，77(2)：257～285

8 Ng H T，Zelle J. Exemplar-based word-sense disambiguation：some recent improvements. In·proc. of the 2nd conference on empirical methods in natural language processing. Somerset，N. J.：Association for Computational Linguistics 1997. 208～213

9 Stanfill C，Waltz D. Toward memory-based reasoning. Communications of the ACM，1986，29(12)·1213～1228

10 Stolcke A，Linguistic knowledge and empirical methods in speech recognition. AI 40 Magazine，1997，18(4)：25～31

11 Jelinek F Statistical Methods for Speech Recognition. MIT Press，Cambridge，MA，1998

12 Wettschereck D，Aha D W，Mohri T. A review and comparative evaluation of feature weighting methods for lazy learning algorithms：[Technical Report AIC-95-012]. Washington，DC：Naval Research Laboratory，Navy Center for Applied Research in Artificial Intelligence. 1996

13 Zavrel J，Daelemans W. Memory-based learning：usig similarity for smoothing. In：Proc. of 35th annual meeting of the ACL，Madrid. 1997a

14 Zavrel J，Daelemans W，Veenstra J. Resolving PP attachment ambiguities with memory-based learning In：Proc of conf on computational natural language learning. 144. Mack Elison-edited. Madrid. 1997b. 136～144

15 Zelle J. Using Inductive Logic Programming to Automate the Construction of Natural Language Parsers：[Ph. D]. The University of Texas at Austin. 1995