计算机科学 1999Vol. 26№. 9

一种基于 CBR 的知识库求精模式*)

Knowledge-Base Refining Schema via CBR

郭宏蘧 李 未

(北京航空航天大学计算机系 软件开发环境国家重点实验室 北京 100083)

Abstract This paper proposed some principles to refine knowledgs base. Underlying the background of machine translation, a knowledge-base(KB) refinement schema is provided, in which, the importance degree of errors serves as the refining indicators and most serious errors are solved first; and the effectiveness and acceptance of KB refinement is determined based on system performance and the KB convergent gain alternation. The error identification strategy via case-based reasoning is presented. Generalizing and specializing operations based on focus revision are presented as well.

Keywords Artificial intelligent, Machine translation, Knowledge base, Case-based reasoning

由于领域知识以及人们的认识进程具有进化的特性,领域模型总是不完备的,为了增强基于知识系统的自适应能力和可靠性,知识库求精已成为机译系统等基于知识系统实用化的必经阶段。本文给出了一些知识库求精原则;结合机器翻译知识库建构维护的实际需求,提出了一种基于 CBR(Case-based Reasoning)的知识库求精模式。该求精模式以提高系统有效性为核心,以错误严重性为指示器,择重优先求精,依据系统有效性和知识库收敛粒度变化衡量求精操作的可接受性;在知识求精中,我们采用基于案例的错误辨识策略,定义了句子贴近度计算以获取最贴近的标准样例;并采用聚焦修正方式,对相关知识进行最必要的概化和特化求精,使知识库逐渐收敛于全真语句集。这些研究不仅为机译系统的知识求精提供了有力支持,也适用于其他领域的知识库求精。

1. 知识求精策略

领域知识模型化是一个逐步收敛于领域全真语句集的逼近过程[1]。我们针对领域知识以及人们的认识进程的进化特性,结合提高系统自适应能力和可靠性的实际求精需求[2],给出了如下一些知识库求精原则。

(1)**有效性原则** 现有的大部分知识求精工具[3~5]只关注改进知识库的学习策略,极少深入考虑所求精知识库的有效性。我们认为知识求精的目的是提高系统的有效性,其重点应放在专家系统的最终有效性上,而不是放在求精过程的学习能力上,即每次求精后,知识库应更逼近于领域理论的全真语句集,专家系统应更加有效。

(2)**择重优先求解原则** 由于系统性能错误的性质不同,知识库求精应以错误的严重性为指示器,优先求解最严重的错误。不同错误对专家系统的整体有效性的影响不同,错误数目减少并不总是意味着系统性能有所提高,因此,系统性能不可用执行测试例库时所检测出的错误数目简单地加以衡量,对专家系统中涉及的错误严重性加以分类是十分必要的。这种分类依赖于具体应用领域,与错误类型和所涉及的信息元素相关。例如,在语言知识求精中,我们以标准结论为参照点,将机译系统生成的结论分为真肯定、真否定、假肯定、假否定。

定义 1 ∀ 句子 S_i,机译系统处理 S_i 时产生的假设集为 H_i,S_i 的正确假设集为 H_i^*,设 h 为一假设,则:①若 $h\in H_i\cap H_i^*$,h 为真肯定;②若 $h\in H_i\cap \overline{H_i^*}$,h 为真否定;③若 $h\in H_i$,$h\in H_i^*$,h 为假肯定;④若 $h\in \overline{H_i}$,$h\in H_i^*$,h 为假否定。

*) 本文研究得到国家自然科学基金(项目号 69433030F)和攀登计划基金资助。

假肯定和假否定是受反驳结论，是知识库不完备或知识受反驳所导致的。h 为假肯定意味着 h 在不应出现的场合出现了，这表明 h 的获取过易，应强化相关知识。h 为假否定意味着 h 在本应出现的场合中没有出现，这表明 h 的获取过难，应弱化相关知识。机器翻译中的假肯定错误比假否定错误危害性更大，因此，知识求精时同一处理层上的假肯定错误的求解优先级最高。

(3)**最小变化原则** 修正每个错误时，只检测和修改导致此结论的最小知识集，其余知识不变。

(4)**容错原则** 在确保知识库的净有效收敛粒度基础上，允许求精后生成较小的新错误。求精后的知识收敛粒度变化可基于如下启发信息加以判断：

· 若一次求精消解了 p 个假肯定，引发了 n 个假否定，则当 $p \geqslant n$ 时，知识收敛粒度不会下降；

· 若一次求精消解了 n 个假否定，引发了 p 个假肯定，则当 $n \geqslant (p \times \alpha)$ 时，知识收敛粒度不会下降，即假肯定的消解弥补了所引发的假否定的负面影响。α 为假肯定与假否定之间的严重性相关常量，其取值与应用领域相关。

由于求精操作的可接受性不单纯依赖于所消解的错误数目，而主要取决于错误性质及其与系统性能的相关性。因此，上述启发信息也有助于判断知识求精的可接受性。当然，领域专家将最终决定接受、拒绝还是重修正已有知识求精操作。

2. 基于 CBR 的知识库求精模式

2.1 总体结构

由于机器翻译知识库的建立、维护是一项异常艰难耗时的任务，已成为制约机器翻译技术走向实用的"瓶颈"。因此，基于上述求精原则，我们采用组合 CBR 和知识求精技术的新体系结构，构造了汉英双向翻译系统 CETRAN[6] 的知识库求精系统 KBRS。CETRAN 的知识库具有不完善性，但接近正确状态，并且是自协调的(即在句子处理中，不会出现属性赋值冲突)。图 1 给出了由两个阶段组成的机译知识库求精框架。第一阶段运用现有知识库对测试句库进行翻译，获取系统可靠性和效率，发掘系统输出译文与标准译文之间的矛盾。第二阶段利用标准样例库识别深层错误，定位应对错误负责的知识库相关部分，并进行理论修正，检测所做求精对系统性能的影响以确定求精的有效性。如果精确性低于阈值，则通过相关例句或专家指导进一步加以修正，直至用户对修正后的知识库的精确性感到满意。这种循环交互模式为专家提供了改进知识模型的第二次机会。

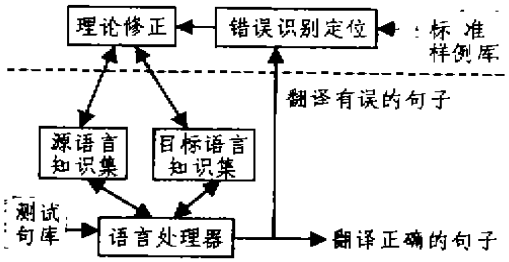

图 1 CETRAN 的机译知识库求精框架

基于择重优先求解原则，我们将知识求精分为消解假肯定和消解假否定两个子阶段，顺次加以激活处理相应一类具体错误。每个子阶段均分为错误识别、错误定位、理论修正三个步骤。由于求解一个错误后，可能又引发一些小错，因此，在每个子阶段开始前，KBRS 将启动错误识别模块，精确确定所剩错误。通常，一个求精系统总会或多或少地进行一些从领域角度看没有什么意义的修正。虽然这些修正会使系统维持一定的性能粒度，但为了确保知识库所含知识的完整性，它们不宜被接受。因此，每次求精后，系统开发专家必须审核所提出的全部修正，原则上只接受从领域角度看具有意义的知识求精。

2.2 基于 CBR 的错误识别定位

在知识求精中，KBRS 采用 CBR 方法[7]，利用标准样例和少量启发性知识，对系统所生成的假设集进行深层错误辨识，确定受反驳假设集以及每个受反驳假设的错误类型、求解级别。KBRS 所利用的标准样例库由系统正确处理的具有代表性的句子组成，以句型为索引。错误识别定位模块首先利用句法范畴约束来检测当前句子的句法依存树[8]是否有效。句法依存树中的一个父子单元是有效的当且仅当该范畴模式在句法依存数据库中可以找到。一个句法树是有效的当且仅当它的父子单元均有效。若当前句子的句法依存树是有效的，则以句法依存树和核心动词的特征信息为线索查找样例库，寻找与当前句子贴近的标准样例。通过比较识别当前句子中隐含的错误。为了在错误识别阶段提高搜索效率，系统一般预先对两个句法依存树进行适当的结构规约和属性抽象。结构规约将屏蔽一些不必要的或冗余的部分，属性抽象则用更加抽象的属性替代已有的属性，如将词汇规约为短语类型，将具体词汇抽象为词类。这类规约抽象知识与语言处理系统共享。

为了从系统的标准样例库中挑选出与当前句子最贴近的样例，我们引入了句子贴近度计算。在给出贴近度计算之前，先定义几个概念。

定义 2 翻译单元是一个具有可翻译特性的树，或者是一个隐去一些可翻译子树后仍保持可翻译特性的树，简记为 u。

定义 3 匹配表达式是翻译单元的组合，简记为 e，通常表示成一个依存树。

句子是由匹配表达式确定的，因此，句子贴近度计算以匹配表达式贴近权重为基础、并遵循如下两个启发策略：

(1)共享的翻译单元越大，贴近度越高；这可由翻译单元的大小 $\Gamma(u)$ 计算，即：

$\Gamma(u)$="u 中结点数目"

(2)匹配表达式中的共享翻译单元既为当前句子依存树的一部分，也为样例句子依存树的一部分，因此，共享翻译单元的这两个依存环境相似度(又称外部相似度)越高，则贴近度越高。

为了实现第二个启发策略，我们首先需要确定两个依存环境中各结点间的最佳对应关系。出现多个候选对应结点时，则利用结点(即单词)之间的相似度确定最相似的结点对，结点之间的相似度取值区间为[0,1]。确定最佳匹配后，外部相似度 $\Psi(u,d)$ 则为：

$\Psi(u,d)$=两个依存环境在最佳匹配下各对应结点之间的相似度之和

其中，d 为句法依存树。进而，翻译单元的贴近权重 $\Phi(u,d)$ 为：

$$\Phi(u,d)=\Gamma(u)\times(\Gamma(u)\Psi(u,d))$$

匹配表达式的贴近权重 $\Phi(e,d)$ 为：

$$\Phi(e,d)=\sum_{u\in e}\Phi(u,d)/\Gamma(d)^2$$

因此，若设当前句子的依存树为 d_c，样例句子的依存树为 d_e，e 为 d_c 和 d_e 的最佳匹配表达式，则两个句子之间的贴近度 $\delta(d_c,d_e)$ 为：

$$\delta(d_c,d_e)=\min(\Phi(e,d_c),\Phi(e,d_e))$$

通过贴近度计算定位与当前句子最贴近的样例之后，可对两个句子进行分层比较，以贴近样例的匹配表达式的相关信息集及所用知识为指示器，获取当前句子处理中生成的受反驳假设集，定位应对错误负责的知识库相关部分。

错误具有传播性，时常会出现一错再错的现象。在同一案例中，许多受反驳假设的出现是某些错误假设诱发所至。为此，我们提出一个系统分组方法用于刻画受反驳假设之间的时序相关性和逻辑推理相关性。在句子处理全过程中，独立于任何受反驳假设而自发产生的错误假设称为父假设，由一些受反驳假设参与激活的错误假设称为子假设。一个父假设及其子女构成一个受反驳假设家族。每个父假设错误均是由于知识约束不完全所致；每个子假设错误则既可能是由于知识约束不完全所致，也可能是因为父假设和其他受反驳假设的共现而合作激发的。由知识约束不完全而诱发的错误假设称为内困型受反驳假设，由某些错误假设合作激发的则称为外扰型受反驳假设。具有推理相关性的错误假设最终将被集成到一个受反驳假设家族中。

知识求精主要关注内困型受反驳假设的消解。引发内困型假肯定(设为 H_p)的原因可能是直接生成 H_p 的规则前件过松或生成 H_p 的规则优先级过高。引发内困型假否定(设为 H_n)的原因主要有：(1)知识库中缺少生成 H_n 的知识；(2)知识库中含有生成 H_n 的知识模块 m，但 m 未被访问。这可能是访问知识模块 m 的条件未被满足或在访问知识模块 m 之前终止执行的条件被满足；(3)知识库中含有生成 H_n 的知识模块 m，m 被访问，但 H_n 未被推出。这可能是推演 H_n 的规则前件未被满足或推演 H_n 的规则被满足但规则优先级过低所致。针对上述原因，应对受反驳结论负责的规则集 R_n 定位，送理论修正模块。显然，受反驳假设家族的刻画有利于有的放矢地消除知识的不完备性和不精确性，提高知识求精的有效性。

2.3 理论修正

基于最小变化原则，KBRS 的理论修正模块采用了聚焦修正策略，即每次只修正一个规则，其余规则不变。修正操作的聚焦是实现最小化知识变动的基础。若翻译有误的句子集为 S_n，对受反驳结论负责的规则集为 R。设 R_i 为当前待修正规则，规则修正将只关注与 R_i 相关的句子集 S_i，其中，结论正确的例句子集(简称为正相关集)记为 $S_{i,p}$、结论受反驳的例句子集(简称为负相关集)记为 $S_{i,n}$，即 $S_{i,p}\cup S_{i,n}=S_i$。若正相关集 $S_{i,p}$ 中与 R_i 相关的翻译单元集为 u_p，负相关集 $S_{i,n}$ 中与 R_i 相关的翻译单元集为 u_n，则理论修正部件将针对具体错误类型、原因，调用相应归纳操作作用于 u_p 和 u_n，识别两个集合及其外部依存环境之间的最显著的差异特征，将之做为规则修正的依据，选择、过滤所需的最小假设集，消除冗余和不一致的信息约束，对 R_i 进行条件概化或特化操作。

当 R_i 对 u_n 中的假肯定错误负责时，R_i 必然对 u_n 中的某个元素施加了不必要的属性赋值，故：

(1)若 R_i 规则条件过松，则应通过增加相关元素属性的否定条件，缩小属性的取值范围，增加新属性或降低属性层次等操作强化规则前件；

(2)若约束条件正确，结论有误，则修正结论；

(3)若规则优先级过高，则适度降低优先级。

从而，形成 R_i 的修正 R_i'，当 R_i' 被触发时，u_n 应与 R_i' 无关，u_p 应仍与 R_i' 保持相关性。

当 R_i 对 u_n 中的假否定错误负责时，则：

(1)若 R_i 规则条件过严，则通过删除一些元素，扩充属性的取值范围，提高属性层次，删除一些属性条件等操作弱化规则前件；

(2)若约束条件正确、结论有误，则修正结论；

(3)若规则优先级过低，则适度提高优先级。

从而，形成 R_i 的修正 R_i'，R_i 被触发时，u_n 和 u_p 均应与 R_i' 保持相关性。

若执行上述相关修正操作后 u_n 中仍有出错的翻译单元，则必须从正、负相关集中提取新规则，两个集合的显著特征充当新规则的前件，理论求精模块提供新规则的结论。当出错翻译单元归属不同的句法结构时，可调用归纳操作区分各类句法结构；当归属同一句法结构时，可通过概化正例引入通用规则模式，依据所有出错翻译单元的相关特征进行初始化。新引入的规则同原有正确的、已修正的规则集合并，组成最终的目标知识库。修正后的知识库要经过原有测试样例和新典型例句(即没有参与修正的、与已进行的知识修正无关的句子)的有效性检测。测试库将依据运行结果被划分为求解正确和求解有误的两个互补子集，以利于用精确性衡量原有知识与已修正知识之间的距离，进而精确地衡量系统性能，完成有效性证实。

2.4 机译系统性能的改进

汉英双向翻译系统 CETRAN 现有三千多条规则，由二十多个规则模块组成。KBRS 作为 CETRAN 的知识求精原型系统，以句子集的系统测试结果为基础。如果测试句库没有覆盖具有代表性的语言问题，则 KBRS 对知识库所做的测试将具有片面性，对系统所做的性能改进也必然只倾向于用以支持求精的几类测试案例上，因此，CETRAN 的测试句库是由专家精心挑选的典型例句组成的，覆盖了相关语言的基本语言现象。这样，随着测试库中句子的正确求解，KBRS 可有效地提高 CETRAN 的系统性能。通过对 KBRS 已修正的假肯定、假肯定数目进行综合度量，可以较直观地揭示 CETRAN 性能的变化。

在使用 KBRS 之前，CETRAN 运行一个含有 100 个典型句子组成的测试子集时，共出现假肯定 79 个，假否定 73 个。KBRS 运行后，在 KBRS 的两个求精子阶段的运行结果如表 1 所示(百分比已近似取整)。

假肯定求精阶段减少了 53%的假肯定错误，增加了 7%的假否定错误，所求解的假肯定错误数目大于新生成的假否定错误的数目，这说明该求精阶段的整体效应是积极有效的。假否定求精阶段没有引起新的假肯定错误，而且减少了 21%的假否定错误。由表 1 可知，与初始的 CETRAN 相比，使用 KBRS 后，假肯定错误数目减少了 53%，假否定错误的数目减少了 15%，这表明知识求精后的 CETRAN 的整体性能有所提高。与此同时，由表 1 还可看出，各求精阶段在 CETRAN 性能改进方面所起的作用略有不同：CETRAN 在假肯定求精阶段获得较大的性能改进；在假否定求精阶段的性能变化则相对较弱一些。这是遵循先假肯定后假否定的求精优先权的合理结果。

表 1 CETRAN 在 KBRS 运行阶段的性能变化

	初始	假肯定求精阶段之后	假否定求精阶段之后
假肯定错误数目	79	37(−53%)	37(+0%)
假否定错误数目	73	78(+7%)	62(−21%)

结语 知识库求精是机译系统和其他基于知识的系统实用化的瓶颈。本文给出的知识求精原则具有一定的通用性，为知识库求精系统的构建奠定了基础。现有的大部分知识求精工具只关注改进知识库的学习策略，极少深入考虑所求精知识库的有效性。机译知识求精系统 KBRS 则将知识求精的重点放在提高系统有效性上，在知识求精中，将受反驳结论分为假肯定、假否定两类，以错误严重性为指示器，优先求解最严重的错误；以系统有效性和知识库收敛粒度变化为依据衡量求精操作的有效性、可接受性。本文提出的句子贴近度计算，能够准确获取最贴近的标准样例，为基于 CRB 的错误识别定位提供了有力支持。KBRS 的聚焦修正策略能够有效地对相关知识进行最必要的概化和特化修正。这些研究不仅为 CETRAN 的知识库建构和维护提供了有力支持，也适用于其他领域的知识求精。

(参考文献共 8 篇，略)