

# 基于描述逻辑本体推理的语义级中文校对方法<sup>①</sup>

姜 赢, 庄润钹, 吴焯凡, 朱玲萱

(北京师范大学珠海分校 管理学院, 珠海 519087)

**摘 要:** 近年来中文校对技术虽然在字词级和语法级层面取得了较好效果, 但是对于语义级层面研究相对薄弱, 目前相关研究都有一定局限性. 提出利用本体技术将中文文本中的语义内容提取出来转换为结构化本体, 再与正确的领域背景本体库融合, 通过描述逻辑推理机来判断提取的语义内容的逻辑一致性, 并将检测出的逻辑一致性错误映射为中文语义错误. 此方法在政治敏感错误等领域进行了测试, 与其他相关研究相比, 具有语义查错的透明性、语义模型的完整性、语义推理的智能型等优势.

**关键词:** 中文校对; 语义校对; 本体推理; 描述逻辑; 推理机

## Semantic Level Chinese Proofreading Method Based on Description Logics Ontology Reasoning

JIANG Ying, ZHUANG Run-Bo, WU Ye-Fan, ZHU Ling-Xuan

(School of Management, Beijing Normal University(Zhuhai), Zhuhai 519087, China)

**Abstract:** In recent years, the Chinese proofreading technology has achieved good results at the word-level and the grammar-level, while it's relatively weak at the semantic level. This paper provides a new method of Chinese proofreading powered by the ontology technology. The semantic contents are firstly extracted from the Chinese texts and transformed into some kind of structured ontology, which is combined with the correct background ontology. The logical consistency of the extracted semantic content is determined by description logic reasoning machine, with the detected logical consistency error mapped into some Chinese semantic errors. This method is tested in the domain of political sensitive information. Compared with other methods, it has obvious advantages of the transparency of the semantic proofreading, the integrity of the semantic model and the intelligence of the semantic reasoning.

**Key words:** Chinese proofreading; semantic proofreading; ontology reasoning; description logics; reasoning machine

随着计算机技术的不断发展, 报刊、图书等传统媒体的文字录入、编辑、排版、校对和印刷已经电子化了. 包括办公室和家庭在内的各种互联网用户在中文文字处理及校对软件上的需求量也非常大. 现有的中文校对软件(例如, 黑马校对系统、金山 WPS)可以根据中文词法关系和语法结构, 自动分析中文文本中的词句并将疑问处标红, 可以一次滤掉大量的文字录入错误, 包括丢字、多字、错字和语句不通等, 特别适合二、三连校中的而校及终校把关<sup>[1]</sup>. 但是目前中文校对软件还不能完成替代人工校对, 其中一个重要原因

在于, 虽然其字词级和语法级层面取得了较好效果, 但是语义级层面中文校对技术研究相对薄弱<sup>[2]</sup>. 例如, “马英九在台湾的职务是总统”这个句子既没有字词错误, 也没有语法错误, 但是其语义错误的原因在于: “由于台湾不是一个国家, 所以它不能有总统”, 需要在“总统”前添加“所谓”, 但这样的背景知识只有专业校对人员进行人工判断, 而中文校对软件通过简单的关键词过滤是无法准确处理的: 如果只是简单判断同时包含“马英九”和“总统”的句子为错误语句, 将导致“马英九给美国总统致电”这样的正确句子错判.

<sup>①</sup> 基金项目: 国家社会科学基金青年项目(14CTQ041)

收稿时间: 2016-07-10; 收到修改稿时间: 2016-08-31 [doi: 10.15888/j.cnki.csa.005680]

## 1 语义错误与语义校对的定义

由于英文文本的词与词之间是以空格符为分隔符的,所以英文自动校对基本以词的校对为核心,是在非词错误和真词错误这两个层次上进行的.研究发现英文文本中非词错误占60%,真词错误占40%<sup>[3]</sup>.其中,非词错误即字串不是词典中存在的;真词错误即字串是词典中存在的词<sup>[4]</sup>,但它与上下文搭配不当,引起句法的语义错误,故真词错误也是语义错误的一种.国内也一般认为语义错误是指一些语言错误是体现在语义的层面上,即在字词层面和语法搭配上不存在问题,而是在语义层面上的搭配有误<sup>[2]</sup>,包括一些表达的内容在某个具体的语境中不该出现的搭配(例如,“聪明的手”).文本校对对类型分为字词级、语法级和语义级3类<sup>[5]</sup>.综上所述,本文研究的语义级中文校对是指针对汉语文本中上述语义错误而进行的文本校对,简称中文语义校对.

## 2 国内外研究现状

中文语义校对研究始于2003年,郑逢斌教授在2003年指出:“辨识一个语句的意义首先要了解其中每一个词素或词的语义,积词成句,再了解句子的语义,然后按上下文的语义来理解全篇.让机器能完全理解和表示自然语言当然是一条求之不得的途径,但由于自然语言的复杂性,这条途径困难很大.”<sup>[4]</sup>.2003年罗振生教授指出文本自动校对中的语义错误检查仍相当困难,但不是无从着手<sup>[3]</sup>,张仰森教授于2006年也提出同样观点,并指出在未来这方面仍需深入研究,呼吁加强句法、语义层次的校对策略研究<sup>[6]</sup>.根据各个专家研究内容和思路的不同将国内外语义校正研究划分成以下三类.

### 2.1 模糊语义对比方法(黑箱模型)

2003年郑逢斌等给出了用句子语义骨架表示句子语义的具体方法和表示形式,然后计算文本中语句与知识库中相关知识进行模糊匹配而得出语句的错误程度<sup>[4]</sup>.2010年Kai A. Olsen等也提出通过模糊比较目标句子与大型文本库中相似句子的方法来鉴别文本拼写、语法甚至语义错误的方法<sup>[7]</sup>.以上两种都属于模糊语义对比方法,他们的出发点在于认为完全精确的理解句子语义可行性不高,因此采取绕过精确语义理解的思路,通过建立某种模糊语义模型,将知识库或文本库中的正确句子与目标句子进行模糊匹配来判断语

义错误.这种方法本质上是一种黑箱模型,虽然知道有语义错误,也可以计算语义错误程度值并选择性的纠正错误,但是其局限性在于不知道具体有什么语义错误,语义错误类型是什么,为什么是语义错误,以及为什么纠错之后的是正确的语义.

### 2.2 精确语义匹配方法(白箱模型)

2009年程显毅教授提出基于HNC的中文文本校对系统模型<sup>[5]</sup>.HNC是由中科院声学研究所的黄曾阳先生专门针对汉语的特点而提出来的一种用于自然语言理解的理论<sup>[8]</sup>,该理论由语言概念空间考察自然语言空间,以概念联想脉络为主线,建立一种模拟大脑语言感知过程的自然语言表述模式和计算机理解处理模式.但是由于HNC本身并没有语义推理机制,需要通过穷举57种句式语义来判断语义错误,不适合大规模语义校对应用.此类属于精确语义匹配方法,即使用某种语义知识表示模型提取文本中的语义对象以及语义对象之间的关系,精确判断文本的语义错误类型和错误原因.这是基于白箱模型思路的一种解决语义错误的方法<sup>[5]</sup>.

### 2.3 语义搭配校对方法

2003年罗振声等提出统计和规则相结合的校对策略,既能检查局部语义限制,也能检查长距离的语义搭配<sup>[4]</sup>.2010年张仰森等提出了一种基于《知网》义原搭配的有效自动查错方法<sup>[9]</sup>.2012年张仰森等进一步提出基于知识库的多层级中文文本查错推理模型<sup>[2]</sup>.对于此类研究,罗振声教授明确指出:“这里的语义检查同传统的语义分析是两个完全不同的概念,它并不试图建立句子的语义框架,只是从成分的搭配关系上加以考察”<sup>[4]</sup>.也就是说此类方法只能检查单一的语义搭配错误,而无法处理其他语义错误.

## 3 总体思路

自2004年OWL第1个版本被W3C联盟推荐为国际标准以来,基于描述逻辑的本体一致性推理理论研究非常热门,目前其推理算法研究已相对成熟.OWL本体能够描述各个领域常见语义对象及其语义关系,是解决各个领域语义问题广泛应用的理论模型,被认为是下一代语义网技术基石<sup>[10]</sup>.OWL建立在严格的描述逻辑基础之上,只要语义推理规则制定正确,其推理结果准确度可达100%.利用本体推理技术分析和诊断本体中的逻辑不一致性,最终使之成为逻辑

上一致的本体,成为本体论的一个重要理论任务。

本文提出基于描述逻辑本体推理的中文语义校对方法总体思路如下:利用本体学习技术将中文文本中的语义内容提取出来转换为结构化本体,再与正确的领域背景本体库融合,通过描述逻辑推理机来判断提取的语义内容的逻辑一致性,并将检测出的逻辑一致性错误知识映射为中文语义错误。此方法的基本假设和基本观点如下:

(1) 基于本体论的中文语义校对模型,发挥了本体论在语义描述、语义推理、语义演化等方面的优势,是白箱模型思路的一种彻底解决中文语义错误的根本办法,是中文语义校对技术发展趋势之一。

(2) 随着本体论、语义 Web 和语义技术的不断发展,各种领域本体库将会越来越多,获取方式的开放化和数量的规模化是领域本体库未来发展趋势。即使没有可以通过免费协议获取或商业途径购买的特定领域本体库,也可以通过人工方法或半自动方法自行构建特定领域本体库,以满足不同中文语义校对应用场景的需求,相关工具软件和技术方法也将越来越成熟。

(3) 中文语义提取是中文语义校对的核心步骤之一,但它不需要对中文语句进行完全的语义分析,只需针对特定领域本体库提取关键三元组语义信息用于中文语义查错即可,可以通过不断提高中文语义错误的召回率和准确率来逐步完善中文语义提取算法,这比进行完全的语义分析可行性要高。

(4) 近几年基于描述逻辑的本体一致性推理,其瓶颈在于对海量数据的大规模推理算法优化问题上。而中文语义校对一般用于文字处理软件编辑环境,一般针对的特定领域中文文本数据量也不会大(篇幅较大的中文文本也可先做切割处理),因此对于逻辑一致性验证推理性能要求不高。此方法将中文语义校对问题转换为 OWL 本体一致性验证推理的逻辑问题,充分利用了后者语义查错准确度高的优势,而其性能上的劣势也不是大问题。

## 4 关键技术实现方法

### 4.1 基于本体学习的中文语义提取实现方法

利用自然语言处理和本体学习技术,从非结构化的中文自然语言中提取语义内容,进而转换成基于 RDF 三元组的本体结构化形式,提供给下一步语义查错处理。中文语义内容的提取包括语义对象(类和个体)

和语义关系(语义属性以及语义关联)。具体来说,有以下两种实现方法。

(1) 基于规则的中文本体学习:2006年研发了一个中文本体学习 Protégé 插件:OntoLTCn<sup>[11]</sup>,它将中文文本进行词法、句法分析,通过总结语义模式 XML 规则,使用 XPath 匹配的方法抽取出语义对象和语义关系。第一步,利用 icclas4j 中文词法分析 API 对中文文本进行中文自动分词和自动词性标注。第二步,定义 XML 模式匹配规则(Mappings)。每一个匹配规则都包含两个部分:XPath 模式匹配条件(Conditions)和自动创建本体的操作(Operators)。OntoLTCn 能够通过 XML 模式匹配的方式将这些特征词汇和关联映射到领域本体库中的类、个体和属性等等语义内容。

(2) 基于机器学习的中文本体学习:基于规则的本体学习技术在语义对象提取效果较好,而对于语义关系效果一般。2011和2012年在基于机器学习的语义关系提取这个瓶颈问题上有重大突破。基本思路是,利用 DBpedia 等 LOD 关联数据和语料库作为背景知识库,对于给定语义关系从 LOD 中提取实例并从语料库中抽取中文本特征进行模式学习,再通过模式检索、排序、筛选等步骤,从语料库中匹配文本获得新的语义关系实例,转而再回馈输入 LOD 关联数据,这样形成迭代循环机器学习过程。

### 4.2 基于本体一致性验证推理的中文语义查错实现方法

对于被检测的中文文本采取基于段落扫描缓冲区的逐段处理模式来处理,而不是传统中文语法校正的逐句处理模式。将每个段落中提取的若干 RDF 三元组与正确的领域本体库进行融合,利用选取的本体一致性验证推理规则在推理机中进行推理,两者产生一致性逻辑矛盾即为检测出来的中文语义错误。实验可供选择的推理机包括 Pellet 或 Jena 等等。针对备选本体一致性验证推理规则,可以利用本体修正增量式验证推理机制,即采取基于缓冲区和时间窗口的交互式推理排序方法对进行推理性能进行优化。具体来说,为每条推理规则建立影响度关联机制,在推理缓冲区内调度语义错误所映射的一致性验证逻辑错误,使影响度高的先进性推理运算,影响度低的根据前者运算采取跳过或简化等方法提高效率,完成一轮排序之后,时间窗口向后移动,最终完成所有的一致性验证推理。

以政治敏感语义错误及其语义校对作为例子来阐

述具体实现方法, 主要包括以下两个技术步骤:

(1) 政治敏感信息本体库构建: 构建领域本体库之前, 首先要搜集政治敏感信息. 由于政治类信息的敏感性和特殊性, 难以通过百度搜索等常规途径在国内一般网络上调研相关内容. 例如, 直接在关键字搜索框输入“政治类敏感信息”, 这样是收集不到想要的信息. 那么就需要具备查找信息的一些技巧, 可以委婉一点输入“如何避免政治类敏感信息”或“如何避免政治性错误”. 在涉及政治性问题的用语规范方面, 参考了新华社新闻报道中的禁用词. 另外, 还通过使用谷歌收集完善了相关政治敏感信息.

其次, 在领域专家的协助之下, 使用 OWL 对中文政治敏感信息进行语义建模, 构建相应的 OWL 领域本体库. 对政治敏感信息本体库语义内容进行调研和归纳总结, 在借鉴政治敏感信息监测实验基础之上, 建立了知识分类, 包括“机构”、“人物”, “职务”, 其他一切“国家”级别的称谓和内容, 如“国歌”、“国旗”; 建立的具体的实例涉及党政机构名称, 国家领导人姓名职务, 涉及领土主权、对外关系问题、港澳台问题等; 建立的实例关系主要是政治概念的逻辑关系, 如“总统是”的源代码如图 1 所示, 共和制国家和政治人物之间的逻辑关系, 只有“共和制国家”才有“总统”(Domain 语义约束).

```
<owl:ObjectProperty rdf:about="&Politics;总统是">
  <rdf:type rdf:resource="&owl;InverseFunctionalProperty"/>
  <rdfs:range rdf:resource="&Politics;人物"/>
  <rdfs:domain rdf:resource="&Politics;共和制国家"/>
  <owl:inverseOf rdf:resource="&Politics;担任总统"/>
</owl:ObjectProperty>
```

图 1 “总统是”实例关系本体 OWL 源代码

## (2) 政治敏感错误本体一致性推理

使用 Protégé 4.1 进行实验, 绑定 Pellet 的 OWL 推理机, 并提供一致性验证推理解释功能. 差分算法实验的基准词库可以选用《人民日报》公开的 1998 年 1 月份语料库. 一般来说, 词和某领域的相关度, 与它在基本词库统计表的词频成反比, 而与它在 OntoLTCn 的 XML 模式匹配得到的词频成正比. 例如, 通过 OntoLTCn 中文语义提取文本“马英九在台湾的职务是总统”获得 RDF 三元组“台湾->总统->马英九”. 如图 2 所示, Protégé 能够通过可视化方式展示语义错误的解释: “台湾是(rdf:type)一个地区而不是一个共和制国家, 国家和地区是语义不想交的(DisjointWith 语义约束), 只有“共和制国家”才有“总统”(Domain 语义约束), 所以作为地区的台湾不能有总统马英九(Inconsistency 语

义错误)”.

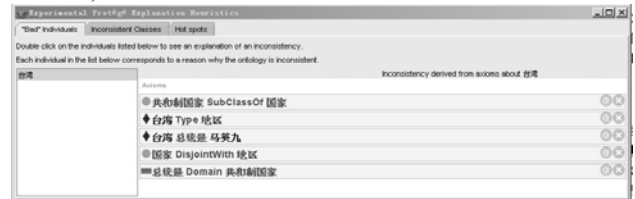


图 2 “马英九在台湾的职务是总统”本体一致性推理校对结果

中文语义纠正首先需要本体推理回溯, 即通过 SPARQL 语句查询知识库中正确语义内容, 例如, “台湾->当局领导人->马英九”. 再将正确语义内容与语义错误对比, 进而自动形成中文语义纠正建议提供给用户选择修改: 将文本中的“总统是”纠正为“当局领导人是”即可.

## 5 计算机系统实现过程

在攻克上述关键技术问题的基础之上, 利用一系列开源软件和自研软件, 使用 Java 语言完成了基于描述逻辑本体推理的语义级中文校对方法的计算机系统实现, 使得计算机能够自动运行中文语义校对.

### 5.1 相关软件介绍

如表 1 所示, 利用了本体处理引擎 Jena 和语法校对 LanguageTool 等开源软件, 极大提高了研发效率. 另外, 对于三元组抽取等关键技术的核心算法, 采取自主研发或二次开发的策略.

表 1 计算机系统实现所使用的相关软件

软件名称	软件类型	主要功能	平台
ictclas4j	二次开发软件	中文分词、词性标注	Java
OntoLTCn	自主研发软件	XML/RDF 三元组抽取 OWL 本体库 CRUD、	Java
Jena	开源软件	描述逻辑推理、推理结果报告	Java
LanguageTool	开源软件	中文语法校对	Java

在进行软件架构设计的时候, 采取面向接口的原则, 避免依赖于某个具体的软件工具. 也就是说, 如果今后能找到替换相应功能的、更好的软件工具, 可以按照相应接口要求进行研发, 嵌入到系统中即可, 其他模块程序不需重写代码. 例如, ictclas4j 可替换为 Stanford NLP, Jena 可替换为 Pellet 等.

### 5.2 具体实现过程

首先, ictclas4j 进行初始化加载 SegTag 类(一次性加载). 利用 SentenceSeg 类的 getSens()方法对中文文

本分句. 利用 SegTag.Split()方法对分句之后的每个句子进行中文自动分词和自动词性标注.

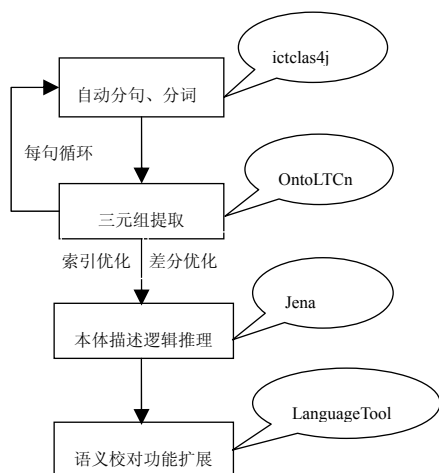


图 3 计算机系统实现过程流程图

其次, 三元组提取采取基于规则的中文本体学习思路, 利用自主研发的 OntoLTCn 的 XML 模式匹配工具进行自动提取. 另外, 加入了一些算法优化: (1)索引优化: 对正确的本体库中的类 Class、实例 Instance、关系 ObjectProperty 和属性 DatatypeProperty 等文字 Literal 建立索引, 每次提取之前扫描一下索引, 至少保证被提取的三元组包含一个本体库中的文字 Literal 才进行下一步推理(如果一个都不包含, 表明句子与本体库领域完全无关, 则可以跳过推理); (2)差分优化: 采取差分算法(chi-square)将词频低而领域相关度高的词抽取出来、将词频高领域相关度高的词剔除(一般来说, 词和某领域的相关度, 与它在基本词库统计表的词频成反比, 而与它在 OntoLTCn 的 XML 模式匹配得到的词频成正比)

然后, 使用 Jena 将正确的本体库读取到 OntModel 中, 并将本体语言设置为 OWL\_DL\_MEM\_RULE\_INF (其中, DL 表示使用 Discription Logics 描述逻辑推理). 将提取的三元组利用 OntModel.addStatement()方法加入到正确的本体库 OntModel 中. 调用 OntModel 的 validate()方法, 获得描述逻辑推理结果 ValidityReport 类对象, 如果它的 isValid()结果是 false 则表明有语义错误. 利用 ValidityReport.getReports()方法获得语义错误列表, 然后使用迭代器循环输出到 LanguageTool 错误报告中.

最后, 在 LanguageTool 传统的语法校对功能上扩展 XML 语义错误规则解析模块, 即在其语法检查工

具 org.languageTool.rules 包中添加类似于针对语语义错误规则 patterns 包, 其中包括(1)XML 语义错误规则数据结构类 SemanticRule 及其(2)解析匹配算法类 PatternRuleHandler. 然后研发继承于语法检查工具 org.languageTool.rules.Rule 类的语义校对规则解析的子类, 覆盖其 getMatches()方法. 然后增加 OWL 推理机调用程序模块、OWL 领域本体库加载解析模块, 使得语义错误在能够映射到 OWL 描述逻辑推理.

## 6 实验测试与分析

### 6.1 实验测试步骤

实验是基于 LanguageTool 中文本语法校对 XML 规则定制方法<sup>[12]</sup>, 主要利用上下文的字、词或词性等特征信息的提取的方法来实现. 在 XML 规则中使用有可能出现语义错误的关键词词进行搭配模型, 来检测一个句子是否存在错误.

首先选择一个适当的生语料库 1(生语料库 1 选取“搜狗实验室的文本分类语料库”), 用所编写的本体库进行推理来检测里面所存在的语义错误. 然后对所检测出的语义错误结果再进行计算机自动判断与统计, 那么出现的结果会有三种情况: ①检测出句子错误的的内容正是预期想要的语义错误内容; ②检测出句子错误的的内容不是预期想要的语义错误内容; ③存在的语义错误没有被检测出. 然后对所收集到的三类数据分别进行计数统计, 计算第一次检测生语料库 1 的正确率 A1 和覆盖率 B1. 接着根据所得到的检测结果修改规则内容, 把修改好的规则拿来重新检测生语料库 1, 统计第二次结果的准确率 A2 和覆盖率 B2 较第一次统计是否有所提高. 换 n 个(本次测试取 n=2, 生语料库 2 选取“新浪微博积极、消极、矛盾数据”)生语料库在进行循环测试, 不断的修改和完善本体库和推理规则, 从而有效提高检错的覆盖率和正确率. 正确率是: ①/(①+②), 覆盖率为①/(①+③).

### 6.2 实验测试结果分析

表 2 实验测试数据结果表

分类	语义错误分类	本体推理规则数量	检测出的数量
通识类	主语和谓语语义搭配不当	5	3
	动词和宾语语义搭配不当	67	23
	主语和宾语语义搭配不当	4	0
	定、状、补与中心语义搭	4	2

配不当			
政治类	涉及港澳台问题	25	12
	涉及领土主权问题	5	4
	涉及党政名称问题	10	5
	涉及前苏联问题	1	1
	涉及党和国家方针问题	2	1
	涉及民族宗教问题	14	9
	涉及国际组织提法	13	8
	涉及法律问题	6	2
	涉及重要人物史实问题	4	1
	总计		161

将政治类和通识类语义错误利用本体推理分别对生语料库 1 和生语料库 2 进行检测, 其中由于生语料库 2 的数据内容过于庞大, 于是在其中随机抽取一部分内容, 其数据量在 175MB 的内容来进行检测. 如表 2 所示, 第一次检测生语料库 1 时, 通识类和政治类所得到的正确率都比较低, 主要的原因是第一次编写本体库时, 没有意识到一些符合推理错误条件的文本内容但是正确句子的情况, 导致检测的结果中存在较多的误检结果. 通过具体实例来实现修改完善本体库, 将一些误检结果最大程度的排除. 于是在修改后的本体库第二次检测生语料库 1, 通识类与政治类的正确率得到了明显的提升, 都在 78% 以上. 生语料库 2 是“新浪微博积极、消极、矛盾微博数据”. 由于在微博上人们的言论相对自由, 故会存在比较多的语义级错误. 并且是基于生语料库 1 的本体库已较为完善, 第一次检测生语料库 2 两类都取得了较高的正确率. 但通过计算机进一步的自动比对, 还是存在一些误检的结果, 通过修改本体库正确率都得到了一定的提升, 但是无法做到 100% 正确率, 原因是一些检测的语句需要判断具体语境的语义错误, 而所选择的语料库内容中一些是不符合该语境于是无法排除该校对错误的检测结果, 这也是实验的局限性所在.

表 3 分类实验测试正确率(%)统计表

生语料	通识类		政治类	
	第一次测试	第二次测试	第一次测试	第二次测试
生语料库 1	59.7	78.3	62.3	78.4
生语料库 2	82.3	87.9	86.4	89.1

## 7 结语

综上所述, 与其他现有方法相比, 此方法在以下三个方面具有明显的优势: (1)语义查错的透明性: 此方法利用 OWL 本体提取文本中的语义对象以及语义

对象之间的关系, 精确判断文本的语义错误类型、错误原因以及如何纠错, 不存在模糊性和不可知性. 这种白箱模型的透明性是模糊语义对比方法无法比拟的. (2)语义模型的完整性: 选择的 OWL 本体, 是一个体系结构非常完整的语义知识表示和语义推理模型. OWL 本体能够描述包含语义搭配校对在内的各种常见语义对象及其关系(语义搭配校对可以使用“对象定义域/值域”建立映射). (3)语义推理的智能性: 语义推理基于描述逻辑, 它内置一致性逻辑推理机制, 其本体推理算法相对成熟, 可以直接利用 Pellet 或 Racer 等描述逻辑推理机进行智能自动语义查错, 无需额外穷举句式或者建立模糊语义骨架. 未来拟将此方法整合到现有的面向词法和语法的中文校对软件中, 使其提供语义级中文校对增值服务; 另外, 还将进一步测试此方法在其他不同领域中的应用情况并加以改进.

## 参考文献

- 1 吴明. 最新版黑马校对软件在新闻出版单位使用. <http://data.chinaxwcb.com/epaper/2011/2011-06-20/11589.html>. [2015-12-18].
- 2 吴林, 张仰森. 基于知识库的多层级中文文本查错推理模型. 计算机工程, 2012, 20: 21-25.
- 3 骆卫华, 罗振声, 龚小谨. 中文文本自动校对的语义级查错研究. 计算机工程与应用, 2003, 12: 115-118.
- 4 郑逢斌, 陈志国, 姜保庆等. 语义校对系统中的句子语义骨架模糊匹配算法. 电子学报, 2003, 8: 1130-1140.
- 5 程显毅, 孙萍, 朱倩. 基于 HNC 的中文文本校对系统模型的研究. 微电子学与计算机, 2009, 10: 49-52.
- 6 张仰森, 俞士汶. 文本自动校对技术研究综述. 计算机应用研究, 2006, 6: 8-12.
- 7 Olsen KA. A smart proofreader for all natural languages: Achieving semantic understanding by majority vote. ISRN Artificial Intelligence, 2012: 1-6.
- 8 史燕, 程显毅, 杨天明, 等. 知网、HNC 和框架网的语义知识表示异同. 广西师范大学学报(自然科学版), 2009, 1: 173-176.
- 9 郭充, 张仰森. 基于《知网》义原搭配的中文文本语义级自动查错研究. 计算机工程与设计, 2010, 17: 3924-3928.
- 10 Cuenca Grau B, Horrocks I, Motik B, et al. OWL 2: The next step for OWL. Journal of Web Semantics, 2008, 6(4): 309-322.
- 11 Jiang Y, Dong H, Xiong H. OntoLTCn: A chinese text oriented semi-auto ontology knowledge discovery tool. Proc. of International Conference on Computer Science and Software Engineering (CSSE 2008) Volume 6. Los Alamitos. IEEE Computer Society. 2008. 18-22.
- 12 姜赢, 曾杰, 林启红, 郭颖珊, 廖文生. LanguageTool 中文本语法校对 XML 规则定制方法. 图书情报工作, 2014, (3): 86-91.