文章编号：1003-0077(2018)10-0001-10

受控自然语言的应用和处理

薛 平

(波音公司 技术研究院,美国 西雅图 98124)

摘 要：自然语言是人类交流最自然的方式。但其复杂性和模糊性常常给有效的交流带来问题。现代社会尤其是当前信息时代面对大量的信息数据,不少工业场景和科研领域以及各种人机交互的应用要求清晰精准、标准化而又较为自然的表达和交流,受控自然语言随着这些需求应运而生。该文讨论受控自然语言及其性质、分类和应用,以及受控自然语言的计算处理方法。该文将以航空工业民用飞机所涉及的英语文本数据为例来阐述受控自然语言在工业场景中的作用和重要性,并且简要讨论受控自然语言更为广泛的意义和价值,涉及其他领域包括当前热门的人工智能等相关的课题。

关键词：受控自然语言;工业语言规范;语言自动检测;知识表达;人机互动;人工智能

中图分类号：TP391 **文献标识码**：A

Controlled Natural Language Applications and Processing

XUE Ping

(Boeing Research and Technology, The Boeing Company, Seattle 98124, America)

Abstract: Natural language is the most natural means for human communication. But its complexity and ambiguity often pose challenges for effective communication. In modern societies, especially during this Information Age, a number of industrial scenarios and scientific areas as well as various scenarios of human-machine interaction require precise but natural information representation and communication. These requirements motivated the concept and development of controlled natural languages (CNLs), which aim to achieve an optimal balance between information precision and naturalness to support effective human-to-human communication and human-machine interaction. This paper discusses CNL, its properties, applications and computational processing. It uses commercial airplane technical documentation as a use case to show the importance of CNL. It also discusses the significance of CNL to other areas such as the area of artificial intelligence.

Keywords: controlled natural language; industry language standards; automatic language checking; knowledge representation; human-machine interation; artificial intelligence

0 引言

自然语言是人类表达和交流最自然的方式。但其复杂性和模糊性也常常给有效的表达和交流带来困难和问题。现代社会,尤其在当前信息时代,有不少领域对相关专业信息的清晰性和精准性有很高的要求,以确保工程技术人员能够准确有效地理解并交流信息。其中有两大类领域。第一类为由于系统和相关程序的复杂性而涉及大量的专业技术信息。这些技术信息必须清晰、精准并且易于专业人员阅读,以确保他们正确地理解其原理和要求,执行相关的程序,维持系统的正常运行。在这些领域中,语言的规范化和标准化至关重要,民用飞机的设计、生产、营运和维修就是个典型的例子。例如,欧洲航空航天和国防工业协会(AeroSpace and Defence Industries Association of Europe)的技术信息标准S1000D,包括 ASD-STE100[1]。第二类涉及人机互动,要求相关的语言既便于人的使用又适合计算机自动分析处理。典型的例子包括人工智能领域中的多种应用,例如,自然语言界面。这些类型的规定和要求推动了受控自然语言的产生和发展。严格定义

收稿日期：2018-07-29 **定稿日期**：2018-08-25

的受控语言有助于人与人的有效交流以及人与机器之间的交互。受控自然语言企图在自然和精准之间找到平衡,在人的需求和机器的需求之间达到某种程度的妥协。受控自然语言使用受限制的自然语言语法规则和词汇,它通常是常规语言的子集是与常规语言重叠的。这给语言解析和处理带来特有的挑战。

本文讨论受控自然语言及其性质、分类和应用,以及计算处理方法。本文的组织结构如下:第一节讨论受控自然语言的类型和属性,同时介绍几个著名的受控英语的实例以及各自的特性。第二节以航空工业民用飞机所涉及的英语文本数据为例来表述受控自然语言在工业中的应用及重要性,并介绍工业界的语言需求和规范。第三节讨论实现简化技术英语计算机软件检测系统的一种方法,以及受控自然语言广泛的意义和价值。第四节将涉及其他领域,包括当前热门的人工智能等相关的课题。第五节总结本文提出的要点,同时简要讨论受控自然语言在语言解析和处理上存在的挑战。

1 受控自然语言的类型和属性

受控自然语言由于它的多样性而没有一个准确且被大家共同接受的严谨的定义。综合起来,受控自然语言可以被定义为[2]:

受控自然语言是基于某种自然语言人为定义的一种语言。在保留了自然语言大部分自然属性的同时,在词汇、句法以及语义上都加上限制条件以控制自然语言的多样性、复杂性和由此可能产生的歧义。

受控自然语言根据目标和用途可分为两大类:

(1) 第一类是受控自然语言的目标在于提高语言表达的清晰性和精准性。

① 提高文本的可读性;

② 改善人与人之间的信息交流与沟通。

(2) 第二类是受控自然语言的目标在于提供自然而直观的形式语言表达式。

① 对该语言进行可靠的机器分析和计算处理;

② 便于人机交流和互动。

世界上有不少类似的语言都属于受控自然语言范畴。其中,仅受控英语就约有100多种[2]。有记载的受控英语已有80多年的历史[3]。由于起源、发展背景以及应用环境和应用目的不同,各种受控自然语言被赋予了不同的名字,例如,“可控英语”(controlled English)、“可处理英语”(processable English)、“简化英语”(simplified English)、“基本英语”(basic English)等。这些语言虽都属受控语言的范畴,但它们各自有不同的特点和性质,在词汇、语法和语义上受限制的方式和程度也各不相同。有些受控语言有较强的表达力,与此同时对歧义的容忍度也相对较宽。有些受控语言被定义得相当严密、准确,犹如形式逻辑和计算机编程语言。以下简要介绍四种著名的受控英语实例。

(1) 卡特彼勒基本英语

卡特彼勒基本英语(Caterpillar Fundamental English,CFE[4])被认为是有史以来第一个基于行业的受控英语,属于上文提到的第一类受控自然语言,其目标是提高语言表达的清晰性和精准性,以改善人与人之间的交流沟通。卡特彼勒是世界著名的工程机械公司,美国财富100家公司之一,主要开发、生产和销售大型工程机械。随着公司产品日益复杂,与顾客进行顺畅的技术交流变得非常关键和重要,但对于不少顾客,英语并非其母语,卡特彼勒大约有两万多种技术文件。卡特彼勒的初衷是希望用CFE来书写技术文件,并通过CFE的语言培训使专业技术人员不用翻译就理解技术文件。最初的CFE的词汇被限制在800～1 000个单词,且规定了每个单词有独一无二的含义。CFE在语法上采用了一系列的限制,包括以下规则[5]:

① 采用陈述句;

② 避免冗长而复杂的句子;

③ 避免一个句子中的题目过多;

④ 避免过多的形容词和名词连用;

⑤ 使用统一的句子结构;

⑥ 避免用复杂的过去和将来时态。

而这些所谓的限制不过是一般的指南而已,很难说是真正的语法规则。因此,CFE的一大弱点是其不可强制执行性。这些规则都比较模糊,仁者见仁,智者见智,用户可根据自己的理解而执行,因此也难以保证技术文件书写的标准化和一致性。卡特彼勒随后开发了卡特彼勒技术英语(Caterpillar Technical English,CTE[6]),在词汇的用法和语义上采用了较为严格且具体的限制,在语法上也增加了清晰的限制规定。更为重要的是,卡特彼勒开发了配套的计算机软件系统自动分析句法和词汇,通过人机互动来进一步加强和有效地执行语法和词汇的规则。虽然最初的CFE仅仅是感性的规则,但其基本理念产生的影响甚为深远。

(2) 简化技术英语

简化技术英语(The Simplified Technical English Specification, ASD-STE)[①]也属于前文提到的第一类受控自然语言，由欧洲航空公司协会(The Association of European Airlines, AEA)发起，联合欧洲航空和国防工业协会(European Aeronautics, Space, Defence and Security Industries, ASD)与美洲航天工业协会(The Aerospace Industries Association of America, AIA)共同研讨开发，于 1986 年推出了第一个版本。英语是航空航天和国防工业技术文件的世界通用语言。ASD-STE 在常规英语的语法和词汇上都加上了限制，其目标是想要通过限制英语语法规则和词汇用法规则来控制技术文档中语言应用的复杂性，减少可能出现的歧义，提高文档表达的清晰度和精准度，以便技术文档的阅读者特别是非英语母语的阅读者能正确理解文件的内容。由于航空系统本身的复杂性，航空技术文件及其内容非常复杂，但航空系统的正确运行和维护关系重大，因此航空技术文件精准的表达和相关人员对技术文件正确的理解至关重要。美国联邦航空管理局(Federal Aviation Administration of the United States, FAA)对航空技术信息特别是航空维修手册以及技术服务公告(service bulletins)等文件的规范管理和审核十分严格。ASD-STE 采用 60 多条语法规则限制，并发布了字典，明确罗列可以使用的英语词汇以及这些词汇相关的语义和用法。这些限制使 ASD-STE 的文本更清晰、精准且易于阅读。ASD-STE 已成为国际航空和国防工业技术文件的书写标准。近年来随着国际化的推进，ASD-STE 也被其他领域，例如，语言服务、专业英语翻译以及学术界所采用。我们将在第二节“受控自然语言在工业场景中的应用及重要性”中更为详细地介绍 ASD-STE 的规则及特性。

(3) Attempto 受控英语

Attempto 受控英语(ACE[7])属于第二类受控自然语言，在语法规则和语义解释上都受到严格的限制，以致它的每个表达式都可无歧义地被自动转换为一阶逻辑的表达式，因此它事实上是一种形式语言。ACE 限制有明显歧义的英语结构，并且通过使用一组无歧义的解释规则对某些可能有歧义的句子进行唯一而无歧义的解释。当仍有多种可能的解释时，ACE 引擎让用户选择其中之一。ACE 有较为丰富的表达力，它不仅允许使用及物动词和不及物动词，还能用双宾语动词。不仅可用简单句，也可用简单句以递归的方式构建复合句。ACE 保留了许多英语的自然属性，又具有形式语言的特性，可作为统一建模语言(UML)的替换语言，描述软件规范；也可替换网络本体语言(OWL)，描述各种信息以及信息之间的关系。因此，近年来 ACE 及其相关工具已用于书写软件规范、定理证明、知识和语义网表示、查询语言和自然语言用户界面等。更详细的讨论，参阅文献[8]。

(4) 计算机可处理英语

计算机可处理英语(简称 CPL[9])也属于第二类受控自然语言，由波音公司研究院研发。其目标是用来支持知识获取收集，以及建立知识库，特别是建立可推理的知识库。CPL 定义的基本句型是：

- 主语＋动词＋补语(＋修饰语)

CPL 规定一个完整的句子必须有主语、动词和补语，而修饰语则是可有可无。可用名词、形容词、介词短语等做修饰语。其语法限制还包括：不允许用代词；名词的指向必须是特指的；简单句子可用连接词“and”连接构成并列句。

CPL 允许的句子可分为三类：基本事实陈述句、问句和推理规则句。问句包括两种基本句型：

- What is NP?
- Is it true that Sentence?

推理规则句的基本句型为：

- IF Sentence 1 THEN *typically* Sentence 2.

在以上的句型中，Sentence 1 和 Sentence 2 都必须是满足以上 CPL 限制规则的句子，可以是带连接词“and”的复合句。CPL 引擎使用启发式规则和辅助的词汇资源(如 WordNet)来排除歧义、限定语义关系。我们将在第四节更为详细地讨论 CPL 的语法解析、语义解释及其主要的应用。

综上所述，任何受控自然语言都是自然语言的受限版本，在语法、词汇和语义上通过明确定义受到限制，使其减少了复杂性和可能产生的歧义，有助于人们的阅读和理解，或便于计算机对其进行处理。第一类和第二类受控自然语言的区别在于目标不同，因此在词汇、语法和语义上受限制的方式和程度不同。需要指出的是，虽然受控自然语言通常在很大程度上保留了原语言的自然性，但其语法规则和

① ASD-STE 是 European Aeronautics, Space, Defence and Security Industries—The Simplified Technical English Specification 的缩写。参看：http://www.trasportilogistica.ittvt.gov.it/PDF/ASD-STE100-ISSUE7.pdf.

词汇用法的限制是人为的，因此受控自然语言并不是自然衍生的语言，而是一种人造语言，一种构建的语言，使用起来也不是完全自然的。

2 受控自然语言在工业场景中的应用及重要性

2.1 工业界的语言特点、需求和规范

现代社会有不少场景要求清晰、精准而又较为自然的表达和交流。民用航空领域是个典型的例子。大型民用飞机可能是当今最为复杂的工程系统之一。从飞机的设计、生产，到飞行营运，到日常的维护，涉及海量的技术信息。以民用飞机日常的运行维修所涉及的技术信息为例，其技术信息有三大特征。

(1) 技术信息量非常巨大

大型民用飞机，如一架波音 B747 或者空客 A380 客机有大约 300 万—600 万个零件。单架飞机维修手册通常平均超过 4 万页。波音公司每年为大约 13 000 架正在服役的飞机共出版约 3 亿页的技术维修资料(包括增补、更新)。如果这些技术资料是纸质的，把它们一页一页叠起来，相当于十万多英尺高[①]，也就是相当于三到四倍于珠穆朗玛峰的高度，如图 1 所示。

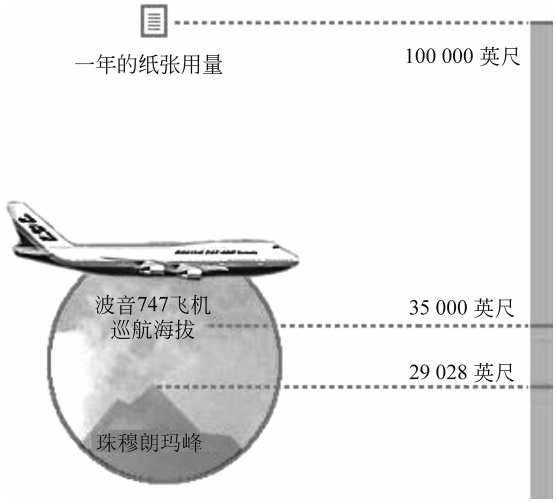

图 1 海量的维修信息

(2) 客户特定的内容和配置极为广泛

即使相同类型的商用飞机也在很多方面会有差异。这常常是源于产品的不断改进以及客户对定制飞机的需求。即使同一公司定制的相同类型的飞机也会如此。飞机这些差异可大可小，可以是用不同的发动机，也可以仅仅是某个断路器位于飞机上不同的位置。这些差异意味着不同的飞机维护程序。因此每架飞机的维修技术资料都可以说是特定的，不能与其他飞机的维修技术资料相混淆。

(3) 十分频繁的技术资料修改和更新

技术信息需要持续和及时地修改更新才能有效。修改和更新的主要原因包括：

① 新的系统配置；

② 工程上的改动；

③ 工程技术程序的改进；

④ 客户的要求和反馈。

因此，波音公司和空客公司都支持频繁的技术文件更新，平均每 90～120 天就会有一次更新。如此庞大的信息数据，加上广泛的配置差异，导致修改和更新相关的部分并妥善管理，确保文件的准确性和一致性成为一项极为繁琐而重大的任务。考虑到这些因素，90 天到 120 天实在不是一个太长的时间。

民用飞机技术资料的复杂性直接反映了飞机系统的复杂性。某种意义上说，飞机的技术资料以及各个部分之间在形式上(包括语言形式)和内容上的关系与飞机系统同样复杂。为了规范编写、出版和管理民用飞机技术信息，美国航空运输协会(ATA)[②]与欧洲航空和国防工业协会分别颁布了信息标准、国际技术出版规范，如 Spec100，iSpec 2200，S1000D，包括数据规范和数据模型。换句话说，飞机技术文件，例如，维修手册虽然是文档的形式，但已经被看成是一种工程规范，要求信息的精准和清晰，有助于准确理解、正确维修，确保系统正常运行和飞行安全[③]。作为技术信息标准 S1000D 的一个组成部分，ASD-STE 简化技术英语成为技术文件的语言国际标准。虽然各国相关的法规并没有硬性要求民用航空的技术文件必须用 ASD-STE 编写，但在实践中，航空公司几乎都在购买飞机的合同中明确要求相关的技术文件，特别是飞机的维修技术手册，包括服务公告(service bulletins)必须使用 ASD-STE 编写。与此同时，政府航空管理部门，如美国联邦航空管理局，对关键的航空技术文件的清晰性、可读性有相当严格的规定和审核程序。事实上，提高技术文件的清晰性、可读性和编写质量也符

① 印刷用纸的厚度通常大约在 0.003 到 0.007 英寸之间。

② ATA 的全称是 Air Transport Association of America。

③ 近年来，ASD 规范 S1000D 以及 ASD-STE 也逐步被用于美国军方陆地、海上和空中合同。但民用航空其他技术文件和工程文件，比如飞行机组操作手册，根据目前的规定不必满足 ASD-STE 规范。

合飞机生产厂家的切身利益。技术文件若不明确清晰，不仅会造成难以想象的事故，而且客户会不断地提出问题，咨询服务并解决这些问题也是不可低估的人力和资金的耗费[①]。

2.2　工业界的语言规范和语言自动检测的需求

第一节提到卡特彼勒公司希望通过 CFE 的语言培训使专业技术人员掌握 CFE，不仅用于编写技术文件，而且能准确理解这些文件的内容[②]。卡特彼勒公司认识到仅仅依靠语言培训还不够，还要有配套的计算机软件协助相关人员使用 CFE。实践证明，计算机软件直接关系到使用受控自然语言的有效性和成败。事实上，绝大多数的受控自然语言都有相应的语言处理系统。受控自然语言虽然基于自然语言，保留了自然语言的很多自然特性，但它毕竟是人为定义的语言，在使用时与自然语言是有冲突的。人们在使用语言时若没有计算机软件检测和辅助，便会不自觉地使用自然语言。而且人工审校受控自然语言文件的合格程度十分有限，效率低，因为使用自然语言是人的本能。

为了更好地遵守、执行国际技术语言规范，ASD-STE 计算机软件检测系统应运而生。例如，波音公司研究院在 20 世纪 80 年代就着手研发了人机交互式语法检查软件系统，被称为波音简化英语检测器（Boeing Simplified English Checker）[③]。并且多年来一直在为完善计算机软件工具而不断努力。这一软件系统自动处理必须符合 ASD-STE 规范的技术文件资料，检查发现不符合 ASD-STE 规则的用法并根据简化技术英语标准提供相关的反馈，以协助文件作者提高文件的清晰度、精准度，以及书写质量。第一节中提到，ASD-STE 包含规则和词典两个组成部分。以下是 ASD-STE 规则的几个例子：

- 规则 1.2：使用 ASD-STE 词典中的 ASD-STE 认可的单词及相关的词性；
  Example：“*test*”只能用作名词 *test*（*n*），不能用作动词 *test*（*v*）。
- 规则 2.1：复合名词不可包括三个以上的名词。
  Example：（*NON-STE*）*university student admission meeting*
- 规则 3.1：不用动词“-ing”的形式，除非动词“-ing”的形式是专用名词或专用名词的组成部分。
  Example：（*NON-STE*）*The indicator is warning the pilot.*
  （*STE*）*The indicator warns the pilot.*
  （*STE*）*The Warning Indicator is on.*
- 规则 3.7：在书写程序或者步骤时，用主动语态而不用被动语态。在描述性语言中尽量不用被动语态。
  Example：（*NON-STE*）*Oil and grease are to be removed with a degreasing agent.*
  （*STE*）*Remove oil and grease with a degreasing agent.*
- 规则 3.8：当词典中有 STE 认可的表述相关动作的动词时，用这一动词来描述相关的动作，而不能用这一词的其他词性。
  Example：（*NON-STE*）*This is an indication of system failure.*
  （*STE*）*This indicates system failure.*

3　ASD-STE 计算机软件检测系统

计算机自动检测 ASD-STE 规则软件系统必须严格遵守 ASD-STE 规则，准确识别任何违反 ASD-STE 规则的词汇、短语和句子。ASD-STE 的规则在常规英语规则的基础上加上了额外的限制。凡是符合常规英语规则且不违背 ASD-STE 规则的语言都是 ASD-STE。因此，ASD-STE 检测软件系统要有实际的用处，不仅能对文本中不符合 ASD-STE 规范的错误做出识别和反馈，还要识别不符合常规的英语语法用法等错误。换言之，ASD-STE 检测系统既需要实现常规英语语法规范，还得实现 ASD-STE 的规范。我们在下文中介绍一种实现这一软件系统的方法。这一方法基于常规英语语法规则和 ASD-STE 规则，详细地表达词汇、语法规则以及语法关系的信息，对输入的语句进行词汇和句法解析。绝大多数语法解析系统的处理方法都是只受理符合语法的短语和句子，只能给符合语法的短语和句子作完整的解析。要识别不符合语法的短语和句子并指出相关的错误，必须延伸

① 据报道，波音和空客的技术文档有关部门每年约耗费 20% 的经费用于回应、解决客户提出的问题。

② 我们把卡特彼勒受控英语通称为 CFE，实际包括 CFE 和 CTE。

③ 参看：https://www.boeing.com/company/key-orgs/licensing/simplified-english-checker.page.

扩展语法解析系统，使其能受理不符合语法的短语和句子，并能够对不符合语法的短语和句子中的错误进行解析。

第一节中提到受控英语是建立在常规英语的基础上，使用受限制的语法规则、词汇和语义，它通常是常规英语的子集，与常规英语重叠。违反英语语法、ASD-STE规范的常见错误包括的错误句型、错误词汇可以看成一个语言的模式集合，以“违规规则”（包括“违规词条”）的形式分别概括每一类错误，汇集起来以定义这个违规集合①。常规英语、ASD-STE、常见的语法及用法错误各自体现的集合分别由三组规则来定义。三者之间的关系示意图如图2所示②。如此延伸扩展的语法解析系统体现了这三组规则的集合。

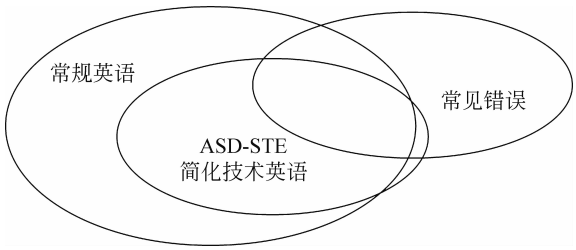

图2 常规英语、ASD-STE和常见错误的集合关系

常规英语规则⇒常规英语

常规英语规则和ASD-STE规则⇒ASD-STE简化技术英语

违规规则⇒常见错误

常规英语语法词法规则定义常规英语。ASD-STE语法词法规则在常规英语的基础上规定了额外的限制，定义简化技术英语。因此凡是没有违反ASD-STE规则、符合常规英语规则的语言现象都在ASD-STE范围内。“违规规则”定义常见错误，即违反ASD-STE规则的语言现象。常见错误包括符合常规英语语法、词法规则但违反ASD-STE规则的语言现象，比如包含三个以上名词的复合名词、被动语态的句子等。图2中违规规则虽然与常规英语规则部分重叠，但扩展了常规英语规则定义的语言范围，使满足违规规则的结构得以编译并生成解析树图。因为所有的违规规则都带有违规特征标志，语法检测算法跟踪记录了使用违规规则在解析树图中的位置，这一信息正是诊断错误的关键。基于以上三组规则，软件系统对输入的语料进行深度解析，对任何违反ASD-STE规则的输入语料给出相应的反馈③。

这一方法采用中心词驱动短语结构语法理论框架[12]。语言模型采用特征结构，详细地表达词汇、语法规则以及语法关系的信息[13]，并采用自下而上的图表解析法对语料进行表达、解析和处理。图3是简单的主谓关系的表达式④。

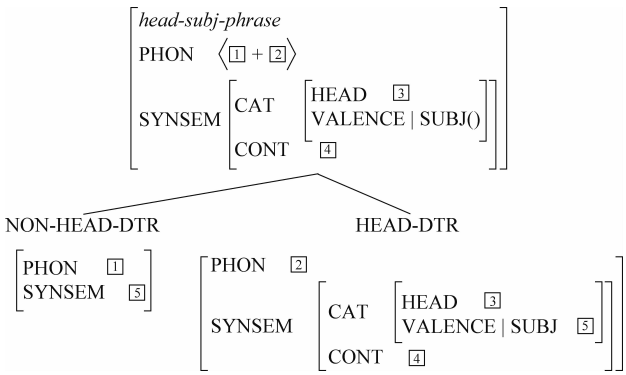

图3 特征结构表达式

特征结构详细地表达了有关词汇、语法规则以及语法关系的信息，比如在以上结构中中心词（即动词）对主语的组合搭配的要求。这里以共享的索引标记5简略表示相互搭配的关系。因此任何主谓不搭配的句子都将在该句的特征结构中表达出来，而使系统能给出详细而准确的反馈。丰富的信息表达加上详细的结构解析使其能够正确识别各种错误的词汇、复合名词及其组成部分、动词特征和形态、词语之间的关系等。

显然，这一基于规则的方法并不是当今流行的自然语言处理的方法。近30年来，绝大多数自然语言处理的研究和应用注重统计模型及相关算法的研究和应用。机器学习，特别是深度学习（deep learning）的方法，被广泛应用在自然语言处理这一传统领域中并取得了不少突破。我们这里采用基于规则方法的主要原因是由文本规则检测这一问题的性质和要求所决定。检测发现不符合规则的用法并根据简化技术英语标准提供相应的反馈是一个精准度要求极高的任务，需要对词汇、句子结构以及语义做深度解析才能准确地指出错误。数据统计的方法能统计出词汇、语串（包括短语），甚至语串之间的相关性，以及句子在日常语料中出现的频率。问题是根据频率特别是根据短语或句子出现的频率难以准确

① 关于“违规规则”的应用，参看文献[10-11]。

② ASD-STE在某些方面可能超越常规英语的允许范围，比如在专业名词的使用上。

③ 波音简化英语检测器软件系统采用了类似的违规规则方法。

④ 有关这一理论框架以及表达式的详细信息，请参阅本文应用的参考资料文献[13]。

地判别这些短语或者句子是否符合语法、是否符合ASD-STE 规则、是否符合规范的语言标准。语言是有层级结构的，句子中的某个组成部分，例如，词或者短语使用得正确与否，常常与句子中的其他部分及其语法功能相互依赖，而且这些相互依赖的句子成分往往不是相互邻接，他们相隔的距离可以是随机任意的，比如上面提到的英语主语、谓语搭配的问题。主谓搭配涉及动词的形态与主语中心词的特征之间的搭配，做主语的名词短语可以任意地复合，比如带多个修饰语和复合从句致使主语的中心词与相搭配的动词相隔甚远。例如，“The university student admission meetings that we attended after the school had started was planned strictly according to the school tradition。”这个句子有主谓搭配的语法错误。主语的中心词“meetings”是复数形式而动词却错用了单数形式。虽然主语、谓语不相邻接，但它们仍需要在单复数的形式上相互一致。

事实上，英语中类似的语言现象非常多。比如，英语疑问句中相关成分移位也常常造成句子成分不邻接。基于数据统计的方法基本上是建立在线性序列基础上，将字或者词等线性符号串作为基本运算单位，语言模型对基本语言单位的邻接关系进行统计和概率度量，不对句子的层级结构及语法功能进行深度的解析，而且也不具有丰富的词汇、语法和语义表达式，因而难以准确地识别句子成分之间的语法关系，特别是远距离的相互依赖关系，难以精准地识别相关的错误以及错误在句子中的具体位置。需要指出的是，虽然我们主张采用这一基于语言规则的方法，但在整个处理过程中不少步骤采用统计的方法有明显优势，特别是在某些预处理和后处理的步骤中，比如词性标记、解析排序、排除歧义及判断习惯的用法倾向等。我们知道语言中有不少现象具有不确定性，取决于场景，例如，输入的句子可能有多种合理的分析和解释。这类具有不确定性的语言现象难以被固定的规则所概括。因此，基于规则的方法与统计方法相结合，在处理的过程中应用了大数据及其相关的分析方法，这不仅对严格的语法规则以及违反词汇、语法、语义和风格规则的错误能够进行准确、清晰的识别，而且对习惯的用法倾向和具有不确定性的语言现象也能做出相应的判断和恰当的反馈。这些功能在文件检测、修订中有多方面的应用。

4　其他领域的应用及意义

事实上，受控自然语言除工业技术文件的编写出版的应用外，还有非常广泛的应用。工业技术文件使用受控自然语言仅仅涉及第一类受控自然语言。其主要目标在于加强语言表达的清晰性和精准性，以提高文本的可读性，改善人与人之间的交流沟通。下面我们简要讨论第二类受控自然语言的应用，其应用与一系列的人工智能的应用相关。第二类受控自然语言被定义得相当严密、准确，提供了一种自然而直观的语言表达式，既能实现受控语言自动机器分析和处理，也便于人机交流和互动。这就为很多应用提供了可能性，也为很多人工智能的应用提供了自然的交互手段。物联网各种器件的人机自然语言交互、机器人或无人飞行器的自然语言控制，以及其他系统的自然语言界面等，都是大家熟悉的例子。这些领域都有特定的场景，其常用的指令是有限的，而且大部分指令不要求复杂而多样化的语言形式，因为特定的场景限定了可能的内容，语义也相对直接而简单明了，但这些应用对语义的解释都要求比较精准。例如，亚马逊的智能音箱 Echo，它通过语音交互除播放音乐、新闻外，还能控制智能家居设备（例如，智能灯、空调、电视等）、叫出租车、从亚马逊网站上订购商品等。Echo 似乎能理解自然语言，事实上它只是执行十分有限的语言指令。这些基本指令语言都可以说是一组受控自然语言。采用受控自然语言是为了方便用户，对用户来说熟悉和使用这些指令十分轻松自如，因为这些语言指令本身就是他们熟悉的语言的一部分。由于它的有限性，使其处理起来也相对简单，这里就不一一详述了。

下面我们简略讨论受控自然语言在知识表达和机器推理中的应用。知识表达和机器推理是人工智能领域十分重要的课题。近年来随着计算能力的迅速增加（参见摩尔定律），人工智能的技术和应用蓬勃发展，出现了一系列的技术进步，其应用深入到广泛的工业自动化领域和日常生活的各个方面。但是，这些技术和应用以及被认为是人工智能的现有系统，绝大多数都不过属于“弱人工智能”（weak artificial intelligence）的范畴，比如 Siri 和 Netflix 就是这类技术产品的典型例子。这些系统都是专注于一项特定的任务，不具有类似人的认知分析能力，及所谓“人工通用智能”（artificial general intelli-

gence),不能根据情况进行推理而解决问题。推理是人类智慧的核心,人工智能进一步发展要求我们对人的智慧、知识以及认知能力做更深入的研究。任何智能都与知识有关,任何智能系统都离不开某种形式的知识集合的支持。知识获取、表达以及建立知识库,特别是建立可推理的知识库,仍然是人工智能的进一步发展中的主要瓶颈。知识的汇集和表达是建立自动推理能力的基础,自然语言当然是人类最具有表现力也最为自然的语言形式,但其复杂性使计算机的自动处理十分困难。大型的知识库,特别是能支持智能系统的知识库,比如语义网(Semantic Web)和知识图(Ontology or knowledge graph)通常是用形式语言(如RDF, DAML+OIL)描述定义的。形式语言有明确定义的语法和无歧义的语义,但是构建特定领域知识库的领域专家常常对这些形式语言并不熟悉,而且形式语言的表达能力很有限。特定领域中的知识概念用自然语言表达十分顺当,但却常常难以用形式语言表达,以致造成理解上的距离甚至差错。正如相关文献中已经提到[14-16],受控自然语言是弥合自然语言和形式语言之间差距的一种方式。理想的受控自然语言应该具有以下三个特征:①具有明确、精准定义的语法和语义,能够被直接转换成某种逻辑表达式,易于计算机自动处理;②基于某种自然语言,语言自然而直观,易于人类理解和表达;③具有足够的表达能力,能够描述表达相关应用领域中的概念和关系。

以上提到的受控英语CPL[9]应用受控自然语言表达知识,建立知识库。与其他语言以及相应的知识库系统相似,CPL配有一个推理引擎作为系统的一部分。CPL计算处理包括三个主要步骤:①解析输入的CPL句子;②将解析的结果转换成一种逻辑表达式;③将逻辑表达式转换成知识推理引擎系统KM的语句①。KM是具有一阶语义基于框架理论的语言,其推理引擎使用了辅助的词汇资源和相关的工具来限定语义关系排除歧义,以取得对输入的语句进行正确的解释。KM系统并没有采用CPL作为推理语言。事实上,受控自然语言可用于整个处理过程,包括推理的步骤。国际技术联盟(International Technology Alliance, ITA)的受控英语ITA-CE就是个例子②。ITA-CE定义了一组常用而简单的句型来表示关于实体存在、属性和关系的命题。同时还应用逻辑规则句和理由陈述句[16]。因此,ITA-CE不仅用受控英语作为表达知识的语言,同时用受控英语作为逻辑语言表达各种事物间的逻辑关系,并且用受控英语作为推理语言。ITA的运作环境是个动态的环境,不断面对大量的结构化和非结构化信息数据。环境中的成员包括人员和机器,人员只是整个网络的一部分,团队之间和跨领域的合作需要不断获取知识、交流信息。ITA-CE的主要目标是支持跨领域的团队之间的信息交换。团队成员通过人机互动获取知识,共享相互之间的信息。关于ITA-CE具体的应用,参考文献[18]。

5 结语

受控自然语言随着社会的工业化和信息化的发展应运而生。工业化和信息化的高度发展带来了一系列的场景,要求既精准而又相对自然的信息表达和交流。ASD-STE以及类似的受控自然语言旨在加强语言表达的清晰性和精准性,以促进人与人之间有效的交流。CPL以及同类的受控自然语言的目的在于实现直观而精准的形式表达式,以便支持人机互动和计算机自动语言分析处理。人与人之间有效的交流和人机互动都要求语言同时具有精准性和自然性。自然语言和形式逻辑语言都不能同时满足这两点要求。受控自然语言是这两者之间的一种折衷。同时要求精准性和自然性是受控自然语言的产生和发展的主要动因。

大型的复杂系统以及其复杂而大量的专业技术信息数据要求信息的标准化和规范化,要求技术信息表达得精准,便于阅读和理解。2.1节中提到的民用航空的技术信息数据事实上与当前大家热论的大数据问题有非常类似的特点。大数据不仅仅是量大,还有其他重要的特征,所谓大数据三要素:数据量大、变化快、多样化③。民用航空的技术信息数据也同样如此:数据量大,动态和持续的修改更新,配置和信息多样化。因此,毫不夸张地说,民用航空的技术信息数据问题是个实实在在的大数据问题。通常大数据的问题仅涉及数据分析,以及有用信息挖

① KM全称the Knowledge Machine,由德克萨斯大学研发。参看文献[17]以及http://www.cs.utexas.edu/~mfkb/km/.

② 国际技术联盟(International Technology Alliance)是由英国国防部和美国陆军研究院共同资助的科研联盟。参看:https://en.wikipedia.org/wiki/NIS-ITA and http://nis-ita.org/.

③ 大数据有三个维度,及所谓三个V:Volume (amount of data)数据量, Velocity (rate of change)变化率,Variety (range of data types and sources)多样性。

掘提取，对信息的精准度没有硬性要求。但民用航空的技术信息不仅涉及信息分析、信息挖掘提取，而且涉及信息数据编写、出版和管理，对内容的精准度有极高的要求，需要满足相关的规范标准。

受控自然语言使用受限制的语法规则和词汇，它通常是常规语言的子集。虽然它是基于某种自然语言，但它也是人为定义的语言。因此，任何受控自然语言的使用与人们使用自然语言的本能在多方面都有冲突，需要额外的努力。正因为如此，受控自然语言的培训必不可少。除此之外，研发和应用受控自然语言软件检测工具十分重要，直接关系到使用受控自然语言的有效性。

本文讨论了基于语法解析来实现 ASD-STE 软件检测系统。这一解析系统包括了定义常见错误“违规规则”，延伸扩展了常规语法解析系统，使其不仅能解析符合语法规则的句子，也能受理并解析常见的违反规则的语言现象。有研究者指出：可能出现的错误范围太广难以预测，用“违规规则”来定义每个可能的错误组合不可行。他们提倡采用放松约束的方法进行解析[19-20]。其他作者却认为虽然不可能用“违规规则”的方法概括所有可能出现的错误，但“违规规则”的方法能针对常见错误提供更为精确的反馈，而且能够覆盖常见的错误类型[11,21]①。事实上，两种方法可以并用。需要分析不同的语言现象和错误的类型，根据各类语言现象的特征，可分别使用“违规规则”或者“放松约束”的方法。

常规语法规则再加上“违规规则”自然会增加语法体系的复杂性，增加分析的歧义和多种选择。而且，一个输入的语句可能同时满足“违规规则”的分析和“常规规则”的分析。换句话说，“违规规则”和“常规规则”可能竞争解析选择。出现这种情况时，就需要大量实际的语料统计数据来帮助解析排序。所以，虽然本文实现 ASD-STE 软件检测系统的基本方法是基于语法解析，大数据和统计方法始终极为重要。

知识和推理是人类智慧的核心。从现在的弱人工智能发展到通用人工智能迫切需要某种形式的知识库的支持。这种知识库采用的语言必须便于领域专家们构建、维护和扩展知识库，又得适合计算机的自动分析处理。受控自然语言是连接人和机器理想的方式。

综上所述，受控自然语言有极为广阔的应用前景，可以总结出三个要点：①作为技术语言标准来规范和改善技术信息的传达和交流；②作为一种直观的知识表达方式，不仅便利专业人员对相关内容的正确阅读理解，而且便于机器自动分析和处理；③作为便利的人机交互的方式和手段，既便于人对设备的操控，又有助于人机互动，最大程度地利用人的智能和机器智能。随着科学技术的不断进步，随着中国进一步走向世界，受控自然语言特别是受控汉语和受控英语的作用和需求会更加明显，这方面的研究和开发也将变得更为重要。

参考文献

[1] ASD (Aerospace and Defence Industries Association of Europe), Simplified Technical English. Specification ASD-STE100, Issue 7[S/OL]. 2017. http://www.asd-ste100.org/.

[2] Kuhn Tobias. A Survey and classification of controlled natural languages [J]. Computational Linguistics, 2014, 40(1): 121-170.

[3] Ogden Charles K. Basic English: a general introduction with rules and grammar [M]. Paul London: Treber and Co., 1930.

[4] Verbeke Charles A. Caterpillar fundamental English [J]. Training and Development Journal, 1973, 27(2): 36-40.

[5] Crabbe Stephen. Controlled languages for technical writing and translation[C]// Proceedings of the 9th Portsmouth Translation Conference, 2009: 48-62.

[6] Hayes Phil, Steve Maxwell, Linda Schmandt. Controlled English advantages for translated and original English documents[C]// Proceedings of the CLAW 1996, 1996: 84-92.

[7] Fuchs Norbert E, Kaarel Kaljurand, Tobias Kuhn. Attempto Controlled English for knowledge representation [C]// Proceedings of the 4th International Summer School 2008 on Reasoning Web, Berlin: Springer, 2008, 5224: 104-124.

[8] Fuchs Norbert E, Kaarel Kaljurand, Gerold Schneider. Attempto controlled English meets the challenges of knowledge representation, reasoning, interoperability and user interfaces[C]// Proceedings of the 19th FLAIRS. AAAI Press, 2006: 664-669.

[9] Clark Peter, Phil Harrison, Thomas Jenkins, et al. Acquiring and using world knowledge using a restricted subset of English[C]// Proceedings of the 19th FLAIRS. AAAI Press, 2005: 506-511.

① 显然，ASD-STE 语法规则检测在原理上与英文教学语法检测批改极为相似。本文讨论的基本方法也适合英文教学语法检测批改的应用。

[10] Schneider David, Kathleen McCoy. Recognizing syntactic errors in the writing of second language learners [C]//Proceedings of the 17th International Conference on Computational Linguistics,1998:1198-1204.

[11] Bender Emily M, Dan Flickinger, Stephan Oepen, et al. Arboretum: using a precision grammar for grammar checking in CALL[C]//Proceedings of the InSTIL/ICALL Symposium. 2004.

[12] Pollard Carl, Ivan A Sag. Head-driven Phrase Structure Grammar[M]. The University of Chicago Press, 1994.

[13] Copestake Ann. Implementing typed feature structure grammars[M]. Stanford University CSLI Publlications,2002.

[14] Schwitter Rolf. Controlled natural languages for knowledge Representation[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics. 2010: 1113-1121.

[15] Clark Peter, Phil Harrison, William R Murray, et al. Naturalness vs. predictability: a Key Debate in controlled languages[C]//Proceedings of Workshop on Controlled Natural Languages 2009, Berlin: Springer,2010,5972:65-81.

[16] Xue Ping, Poteet Steve, Kao Anne, et al. Constructing controlled English for both human usage and machine processing[C]// Proceedings of the 7th International Web Rule Symposium,2013.

[17] Clark Peter, Porter B. KM - the knowledge machine: Users manual[R]. University of Texas at Austin, 1999.

[18] Dave Braines, David Mott, Simon Laws, et al. Controlled English to facilitate human/machine analytical processing[C]// Proceedings of Spie Denfeuse, Security and Sensing, 2013, 8758(3): 875808.

[19] Matthews Clive. Grammar frameworks in intelligent CALL[J]. Calico Journal, 1993, 11 (1): 5-27.

[20] Vandeventer Anne. Creating a grammar checker for Call by constraint relaxation: a feasibility study[J]. RecALL, 2001, 13 (1): 110-120.

[21] Flickenger Dan, Jiye Yu. Toward More Precisionin Correction of Grammatical Errors[C]// Proceedings of the 17th Conference on Computational Natural Language Learning: Shared Task. 2013: 68-73.

薛平(1953－),博士,波音公司研究院(Boeing Research and Technology)资深研究员(2016 年 5 月从波音公司退休)。现担任斯坦福大学语言信息研究中心、北京天学网教育科技股份有限公司研究项目技术顾问、美国国家航空航天局/亚利桑那州立大学航空信息整合及飞行安全项目咨询委员会成员等职。主要研究领域为自然语言处理以及工业领域的应用、计算语言学、理论语言学。

E-mail: p_xue@yahoo.com