理性主义与经验主义相结合的机器翻译研究策略

徐金安
(北京交通大学计算机与信息技术学院 北京 100044)

摘 要 主要介绍了基于规则、基于实例和基于统计等3种主流机器翻译方法,探讨了自然语言处理技术和机器翻译中基于规则的理性主义方法和基于统计的经验主义方法的优缺点,结合机器翻译研究的现状和发展方向,提出了规则和统计相结合的机器翻译方法的基本思路,阐述了词义消歧中的理性主义方法和经验主义方法相结合的发展方向,对机器翻译的发展趋势进行了探讨。

关键词 机器翻译,自然语言处理,计算语言学,理性主义方法,经验主义方法
中图法分类号 H085 **文献标识码** A

Rationalism and Empiricism on the Combination in Machine Translation

XU Jin-an
(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract This paper firstly discussed the development of machine translation, secondly described the advantages and disadvantages of the rationalism and empiricism in natural language processing and machine translation, and thirdly introduced our basic ideas of combination of the rule-based approach and statistical approach. Furthermore, the paper deliberated the tendency of combination of the rationalism and empiricism in word sense disambiguation, finally summarized the developmental tendency of machine translation.

Keywords Machine translation, Natural language processing, Computational linguistics, Rationalist approach, Empiricist approach

1 引言

当前,经济全球化和信息化要求人们努力克服语言障碍,世界上很多国家对翻译的需求急剧增长,潜在市场十分巨大。据美国ABI公司对世界翻译市场的调查结果,2005年翻译市场规模已突破220亿美元[1]。2007年欧盟委员会的官方语言已达23种,翻译人员达2700多人,各机构每年的翻译费用达10多亿欧元[2]。我国在2003年翻译市场的产值为110亿元人民币,2005年200亿,2007年已发展到300亿[3]。随着我国经济的蓬勃发展和国力的提升,汉语热正在世界范围内兴起,翻译服务必会成长壮大为一种新兴的主流文化产业。机器翻译技术在翻译产业中起着十分关键的辅助作用,可以大大减少翻译工作人员的工作量。机器翻译的地位将稳步提升。

机器翻译研究先后经历了20世纪60年代末的低谷期、70年代中期的发展期,80年代开始进入繁荣期和商品化进程,90年代开始进入网络化时期。如今,随着机器翻译研究的深入和发展,机器翻译已经开始进入翻译语言和服务形式的多样化、使用终端小巧化、系统大规模集成化和自动语音翻译系统实用化的新阶段[4]。但是,机器翻译研究依旧面临诸多难题,涉及自然语言处理的不同层面,诸如,单词分割、词性标注、句法结构消歧、语义消歧、知识自动提取等。从自然语言的浅层处理到深层处理,几乎所有的自然语言处理技术和机器翻译方法都要直接面对自然语言问题的复杂性和多变性,现存的理论和方法都还不够完善,发展空间还很大,有待进一步的改进。如何使基于规则的理性主义方法和基于统计的经验主义方法有机地结合,一直以来都是自然语言处理和机器翻译研究的热点问题。

本文第2节介绍机器翻译发展现状;第3节讨论基于规则的理性主义方法和基于统计的经验主义方法的优缺点;第4节结合机器翻译的发展现状,提出规则和统计相结合的机器翻译方法的基本框架和基本思想,分析介绍规则和统计相结合的机器翻译研究策略;第5节探讨语义消歧中的基于规则和统计相结合的发展方向及其部分成果;最后,对机器翻译的发展趋势进行探讨和总结。

2 机器翻译简介

2.1 机器翻译方法

机器翻译作为自然语言处理的一项应用技术,涉及人工智能、数学、语言学、计算语言学、语音识别和语音合成等多种学科和技术,具有综合性、交叉性强的特点,属于国际前沿领域,是目前国际上最具挑战性的研究课题之一。

到稿日期:2010-07-19 返修日期:2010-12-08 本文受中央高校基本科研业务费专项资金(2009JBM027)资助。
徐金安(1970—),男,博士,副教授,主要研究领域为自然语言处理、机器翻译、模式识别和人工智能应用,E-mail:jaxu@bjtu.edu.cn,xja2010@gmail.com。

按照冯志伟老师的说法[5]，机器翻译研究先后经历草创期、萧条期、复苏期和繁荣期等几个阶段，形成了多种机器翻译方法，如：直接翻译方法、转换方法、中间语言方法、基于语言学的方法、基于知识的方法、基于平行语法的方法、基于实例的方法、基于统计的方法等。

在方法论层面，机器翻译系统可以分为基于规则和基于语料库两大类。习惯上人们把直接翻译方法、转换方法、中间语言方法归类于基于规则的翻译方法。基于语料库的方法又可以分为基于记忆的翻译方法、基于实例的翻译方法、基于神经网络的翻译方法和基于统计的翻译方法等。

目前，基于规则的机器翻译方法（Rule-Based Machine Translation，RBMT）、基于实例的机器翻译方法（Example-Based Machine Translation，EBMT）和基于统计的机器翻译方法（Statistical Machine Translation，SMT）占据机器翻译的主流地位。多引擎机器翻译策略（Multi-engine Machine Translation）促进了3种主流机器翻译系统的平衡发展。

基于规则的方法对语言语句的词法、语义和句法结构进行分析、判断和取舍，然后重新排列组合，最后生成等价的目标语言。本文将在第3节对基于规则的方法和基于统计的方法的优缺点作详细的分析和探讨。

基于实例的机器翻译方法的基本思想由日本著名机器翻译专家长尾真提出，其借鉴了外语初学者的学习过程和基本模式。翻译过程是首先将源语言句子分解为一些短语碎片，接着通过类比的方法把这些短语碎片翻译成目标语言的短语碎片，最后再把这些短语碎片构成完整的句子。该方法有多种表现形式和名称，比如基于实例、基于记忆、转换驱动和基于个例的翻译方法等等[6]。

基于实例的机器翻译方法的主要优点在于，首先，翻译结果直接从语料库的实例变换产生，可信度高；其次，避免了规则方法中深层次语言学分析，不需要花费大量的人工调试规则库，翻译的效果随着语料库的增大而逐步提高，易于扩充和维护；再次，翻译算法相对简单，速度快，效率高。主要缺点是：第1，知识的抽象程度低、代表性差，翻译结果很难达到较高的覆盖率，因而在很多情况下都是作为其他系统的补充，作为多翻译引擎中的一个来使用；第2，为了提高覆盖率，在引入深层的句法分析或依存关系分析进行泛化的同时，会产生错误匹配，从而导致泛化和匹配的平衡产生问题；第3，由于词和短语的对齐问题有待进一步提高精度，从而影响了翻译结果的正确率[6]。

早在1949年，美国工程师W. Weaver正式提出了机器翻译问题和类似解码过程的统计机器翻译思想。1990年，Brown等人提出了5种复杂程度递增的数学模型，通常称为噪声信道模型或IBM模型，其中，IBM模型1仅考虑词对词的互译概率，没有考虑源语言和目标语言的句子长度；IBM模型2加入了词的位置变化的概率；IBM模型3加入了一个词译成多个词的概率，考虑了句子长度；IBM模型4考虑了标释集的中心词的概率和其他单词位置的概率；IBM模型5考虑了源语言句子单词间的相对位置的概率。

1998年，王野翊在他的博士论文里提出了一种对于IBM统计翻译模型的改进方法，即基于结构的对位模型。王野翊等以德英口语翻译系统为研究背景，针对德语和英语间语法结构差异较大的问题和数据稀疏问题提出了基于两个层次的对齐模型的策略，即短语之间的粗对齐和短语内词的细对齐的方法，在文法推导方面，采用了基于互信息的词语聚类方法，起到了短语归并和规则学习的作用。该方法既提高了翻译精度；也提高了整个系统的效率使得搜索空间更小；同时缓解了因口语数据缺乏导致的数据稀疏问题[7]。

其后，吴德恺提出了基于反向转换文法和基于概率反向转换文法的对位模型[7]。其基本思路是：对于双语平行语料库中的双语平行句子，可以在正常的上下文无关文法中引入反向产生式，在统一的语法体系之下对双语句子进行同步分析，双语句对同步分析的过程事实上实现了双语句子的结构对齐[7]。反向转换文法较好地考虑了两种语言在句法层次上的约束关系，使双语句子能够在统一的语法体系下进行同步分析。概率反向转换文法考虑了在双语句法分析中的作用，更有利于实现大规模平行语料库的自动对齐和歧义消解，并提高执行速度[7]。

Yamada和Knight等在IBM的统计翻译模型的基础上，受反向转换文法和有限状态转换机思想的启发，提出了一种基于句法结构的统计翻译模型。该方法中，噪声信道模型的输入是一棵句法分析树，信道模型通过对句法分析树的每个节点进行如下分析得到对应语言的翻译句子：①句法树的扁平化处理和节点次序调整；②在每一个节点上随机地插入额外的目标语言的词；③翻译下一个叶节点上的词。该方法属于一种串到树的翻译模型，其主要目的是用于参数训练以获取源语言的字符串和目标语言句子结构之间的对应关系，比如对处理中文句子的字符串和日语的主宾谓结构（SOV结构）等有较好的效果[7]。基于句法分析的翻译模型在近几年的统计翻译方法中得到了广泛关注，典型的有串到树、树到树、树到串翻译模型等。

2002年Och等人借鉴了Papineni的最大熵方法在统计自然语言理解的思路，提出了最大熵统计翻译模型，也有学者称之为对数线性模型。该模型是一种比噪声信道模型更一般化的方法，具有灵活性强、可以随意增加特征，以及可以调整参数等优点，随着其训练方法的逐步完善而被广泛采用[6]。

另外，文献[95]对统计机器翻译做了详细的分析和探讨；文献[96]对基于句法的统计机器翻译研究进行了综合性的分析，阐述了句法结构对统计机器翻译的重要性。

1994年，Frederking提出了一种多翻译引擎的方法，1998年Hogan通过一个简单的实验证明了这种方法确实能够实现比任何一种单一方法都更高的准确率。多引擎机器翻译系统结构可分为系统级多引擎结构和部件级多引擎结构。系统级多引擎结构采用多个完全独立的机器翻译系统进行集成，比如美国CMU等机构研制的Pangloss系统。部件级多引擎结构是在机器翻译系统的多个主要功能模块中分别采用多引擎技术，比如德国教育与研究部（BMBF）资助的Verbmobil系统[6]。近几年，多引擎机器翻译技术发展很快，在一定程度上提高了机器翻译系统的精度、鲁棒性和实用性。

多引擎机器翻译系统的优点是在一定程度上能够互相补充、改善翻译结果、提高翻译精度。而其缺点是系统规模庞大、造价高、不易维护。

2.2 国内外机器翻译现状

目前市场上出售的机器翻译产品绝大部分是基于规则的。早期开发的机器翻译系统主要为国家、国际政府组织机

构和军队服务，典型的例子是加拿大蒙特利尔大学与加拿大联邦政府翻译局联合开发的TAUM—METEO系统，于1976年开始提供天气预报服务。美国空军于1970年研制了Systran，目的是将俄国军事方面的科学技术文献翻译成英语。目前该系统能够处理52种不同国家和民族的语言，提供9个语言包等多种订购服务，支持多种文件格式，服务对象扩大到了个人、家庭、门户网站、跨国公司等。

基于规则的翻译系统还有很多，比如美国的OpenLogos[8]、法国纺织研究所曾经研制的Titus4[9]、美国的Weidner[10]和Paho[11]、德国的Metal[12]和Susy[13]、欧盟的Eurotra[14]、荷兰BSO公司的DLT[15]、日本的Atlas[16]、Mu[17]和ODA计划[18]等。

早期的EBMT系统有：日本的MBT1[19]和MBT2[20]、日本ATR的ETOC和EBMT系统[21]、美国的Pangloss系统中的EBMT子系统[22]等等。日本的国家级重要研究课题“日中中日语言处理技术开发研究”由日本情报通信研究机构主持，2006年初开始启动，为期5年，其中“面向专利等科技文献的中日双向机器翻译系统研发”部分就采用了EBMT[23]方法。

目前，SMT系统产品为数不多，较知名的有Google公司于2005年推出的在线多语言机器翻译系统。近几年，随着统计机器翻译研究的逐步深入，统计模型不断得到改善，翻译精度不断提高，逐步缩小了和基于规则的翻译系统之间的差距。

典型的基于混合策略的多引擎翻译系统有：美国的Pangloss多引擎机器翻译系统、德国的Verbmobil系统、东芝中国研发中心的混合策略机器翻译系统、欧盟于2006年启动的EuroMatrix项目和2009年启动的EuroMatrixPlus项目等等。

早期从事自动语音翻译系统的研究机构有日本电话翻译实验室（ATR）、美国AT&T公司、美国卡内基梅龙大学（CMU）、德国卡尔斯努大学（UKA）、日本NEC公司、德国西门子公司等。最近，便携式自动语音翻译系统产品相继问世，该类产品具有以下特点：面向小型便携式终端、不需要经由服务器进行处理、词典规模小、主要面向旅游口语会话和日常用语等小领域、系统以RBMT为主。日本东芝公司于2009年12月29日宣布适用于手机的日中英三国语言间的自动语音翻译系统开发成功。美国VoxTec公司开发的Phraselator P2便携式自动语音翻译机由美国军方研制，曾经是伊拉克战争美军士兵的军事装备之一，表面经过耐防腐处理，非常坚固。采用直接翻译方法，通过运用大规模电子词典和大量的短语集来实现翻译功能。如今，该产品是美国警方的掌上电子翻译系统，也向民间销售。该产品具有体积小、重量轻、操作方便的特点，能实现英语和42种语言的翻译功能。

我国于1959年成功研制了第一台基于规则的俄汉机译系统之后，机器翻译研究经历了停滞期、复苏期、恢复期和发展阶段。目前，国内从事机器翻译研究的单位主要有中科院计算所、中科院自动化所、中科院软件研究所、清华大学、哈尔滨工业大学、厦门大学、南京大学、东北大学、以及微软亚洲研究院等外资研究机构。在SMT研究领域，我国的科研水平较高，中科院计算所研制的汉英翻译系统在NIST2009[24]测评中获得好成绩，具有国际领先优势。中科院自动化所等单位承担的“863计划”重点课题“多语言自然口语对话系统关键技术研究”项目于2010年1月中旬顺利通过专家组验收[25]。另外，国内有代表性的翻译软件，如金山、译星、华建、英业达等公司的产品都具有一定的技术实力且拥有广大的用户群体。

总体上说，随着计算机软硬件技术和自然语言处理技术的发展，机器翻译研究经历了低谷期、发展期和繁荣期之后，从20世纪80年代进入商品化进程，20世纪90年代进入网络化时代，如今已经进入翻译语言和服务形式的多样化、使用终端小巧化、自动语音翻译系统实用化和系统大规模集成化的实用阶段。

3 基于规则的理性主义方法和基于统计的经验主义方法

冯志伟老师在宗成庆老师的《统计自然语言处理》一书的序言二中[7]，从哲学和方法论的高度系统地描述了自然语言处理中基于规则的理性主义方法和基于统计的经验主义方法之间水火不相容、充满矛盾和对立的历史发展过程，透彻地分析了自然语言处理中基于规则的理性主义方法和基于统计的经验主义方法的优缺点，提倡彼此之间的有机结合，相互取长补短，进行优势互补，为自然语言处理和机器翻译研究指明了方向，在理论研究和实践应用方面都具有重要的指导意义。

3.1 基于规则的理性主义方法

在自然语言处理的发展过程中，以乔姆斯基的形式语言为理论基础的基于规则的理性主义方法，在20世纪60年代末到70年代占据主流地位，基于统计的经验主义方法在此期间则几乎完全受到排斥。乔姆斯基主张以一种公式化、形式化的方法，严格按照一定的规则来描述自然语言的特征，试图以有限的文法规则描述无限的语言现象。这种方法从1970年前后到1990年前后发展很快，且日趋成熟。但是，随着研究的深入，人们也发现：基于规则的语言处理系统性能有限，使用范围受限于某些特定的小领域，在应用于不同领域时，系统的扩展性能很差，往往要求整套规则重写，系统精度很难得到进一步提高。

按照冯志伟老师的说法，基于规则的理性主义方法在自然语言处理中的优点体现在以下几个方面[7]：

1. 规则方法对抽象的语言特征具有很强的形式描述能力和形式生成能力；

2. 规则方法对句子结构以及长距离依存关系的处理能力很强；

3. 规则方法的语言模型结构清晰易懂；

4. 规则方法在本质上无方向性，其语言模型同时适用于分析和生成处理；

5. 规则方法在自然语言处理中具有解决浅层到深层多层次问题的能力；

6. 规则方法与一些高效算法具有兼容性。

在机器翻译研究领域，基于规则的机器翻译系统的优点体现在：

1. 基于规则的机器翻译系统对不同语种在语法上的相似度的依存度很低、处理能力较强；

2. 基于规则的机器翻译系统具有对知识表达的抽象程度高、代表性强的特点；

3. 对文法结构具有很强的保持能力；

4. 对不同语料的覆盖率高；

5. 系统运行占用资源少。

冯志伟老师指出，基于规则的理性主义方法在自然语言处理中的缺点体现在[7]：

1. 使用基于规则的理性主义方法研制的自然语言处理系统的鲁棒性和灵活性都很差；

2. 规则方法需要语言学家对繁杂的语言现象进行大量的分析，语法分析任务繁重，很难用计算机对基于规则的语言模型进行泛化处理；

3. 使用基于规则的理性主义方法研制的自然语言处理系统具有很强的针对性，而且扩展性能差、升级困难；

4. 规则方法不具备统计方法的学习能力，因而领域适应能力较差。

在机器翻译研究领域，基于规则的机器翻译系统的缺点体现在：

1. 基于规则的机器翻译系统的语法分析和生成规则主要由人工编写，规则的主观性强，规则的一致性难以保障；

2. 知识获取难、工作量大；

3. 不利于系统扩充，尤其对非规范的特殊语言现象缺乏相应的处理能力等。

3.2 基于统计的经验主义方法

基于统计的经验主义方法从20世纪90年代开始得到了快速发展，自然语言处理研究也“重新回到经验主义”的发展道路上来，特别是大规模语料库的出现，促使基于统计的经验主义方法成为当前自然语言处理技术的主流。

按照冯志伟老师的说法，基于统计的经验主义方法在自然语言处理中的优点体现在[7]：

1. 统计方法具有良好的数学模型，无指导的学习能力和较强的知识自动获取能力；

2. 统计方法可以比较容易地采用大规模训练语料不断提高系统的性能；

3. 统计方法可以很容易地结合多种规则，通过处理各种各样的约束条件问题，不断改善处理效果；

4. 统计方法善于处理模糊的语言现象。

在机器翻译研究领域，基于统计的机器翻译系统的优点体现在：

1. 目前，网上用于开发统计机器翻译系统的资源丰富，易于实现；

2. 良好的统计翻译模型能够融合更多的句法结构和语义语法信息；

3. 统计机器翻译的翻译性能可通过大规模语料训练进行改善。

冯志伟老师指出，基于统计的经验主义方法在自然语言处理中的缺点是[7]：

1. 使用基于统计的经验主义方法的自然语言处理系统，运行时间和模型参数成比例线性增长，运行效率相对低下，消耗的资源巨大；

2. 统计方法需要大规模语料库进行训练学习，对语料库的质量要求较高；

3. 统计方法容易出现数据稀疏问题，且随着训练数据规模的增大呈线性增长，需要进行平滑处理进行合理解决。

在机器翻译研究领域，基于统计的机器翻译系统的缺点是：

1. 统计机器翻译系统的系统运行消耗的资源巨大，需要大规模双语平行语料库进行训练学习，语料的选择和处理工程量也很巨大；而翻译模型和语言参数的精确性直接依赖于语料的多少，翻译质量的高低主要取决于概率模型的好坏和语料库的质量及其覆盖能力等；

2. 需要解决数据稀疏问题；

3. 统计机器翻译的翻译效果对源语言和目标语言间在语法上的相似度存在一定的依赖性，相似度越小，需要调整的参数越多，系统的运行效率越低，改善系统翻译精度的难度系数越大。

4 规则与统计相结合的机器翻译研究策略

4.1 规则与统计相结合的机器翻译基本思想

规则方法和统计方法在自然语言处理中的优缺点各异，使用不同方法研制的机器翻译系统的性能表现也不尽相同，对译文质量的保障程度深浅不一，翻译精度也还不尽如人意。如何通过规则和统计方法的有机结合，实现相互取长补短的新理论和新方法，是机器翻译研究人员普遍关注的研究热点。

在机器翻译研发过程中，应本着“具体情况具体分析，分门别类地解决问题”的指导方针开展研究工作。从规则方法和统计方法各自的基本原理出发，结合各种机器翻译系统独自的主体特征，采用不同的研究方法和策略，从理论和实践两个方面完善机器翻译理论，实现机器翻译系统的有效改良和升级。

在统计机器翻译系统中，可以采用多种策略融合规则，使规则在统计机器翻译系统中起到关键性辅助作用，根据语言的某些特定语法现象，建立适当的规则体系和模型参数，以在统计机器翻译模型中融合更多的句法结构和语义信息，是当前的统计机器翻译系统改善翻译质量，提高翻译精度的一种思维模式和方法，也是实现规则和统计相结合的机器翻译的一种有效途径。

另一方面，在基于规则的机器翻译系统中，融入多种统计模型，使统计模型在规则型机器翻译系统中起到关键性辅助作用，针对语言中各类复杂文法、非规范文法的具体特征，构建适合其语法特点的统计模型；建立特定文法与统计模型之间的相关约束机制，作为规则与统计方法相结合的切入点。这种思维方式是当前基于规则的机器翻译系统改善翻译质量、扩展领域适应能力的一种行之有效的方法，也是实现规则和统计相结合的机器翻译的另一种有效途径。

以基于转换的规则型机器翻译系统为例，本文所主张的规则和统计相结合的基本思想框架主要表现在以下几个方面：

1. 针对语言中的简单语法和规范语法，采用以规则为主、统计为辅的策略；

2. 针对语言中的非规范文法现象、复杂程度高的语法现象，采用统计为主、规则为辅的策略；

3. 针对语言中不同性质的非规范文法现象、复杂程度高的语法现象，有针对性地构建适合其语法特点的统计模型来解决具体问题；

4. 建立特定文法与统计模型之间的相关约束机制，作为规则和统计方法相结合的切入点。

在对既有的基于规则的机器翻译系统进行领域扩展时，采用上述基本思路和方法，可以有效保留既有的翻译规则；可以避免整套翻译规则重写；可以节省造价、缩短研发周期并提高翻译质量。另一方面，在研发新的基于规则的机器翻译系统时，可以采用上述基本思想，依据文法的复杂程度进行适当的评估和分类，界定简单文法和复杂文法之间以及规范文法和非规范文法之间的区分，建立使用规则或统计模型的粗略标准；针对简单文法、规范文法制定翻译规则；针对复杂文法和非规范文法，有针对性地建立适合其语法特征的统计模型；同时，根据特定文法与统计模型之间的相关约束机制界定规则和统计模型的使用方法。按照这种思路开展研发工作，同样可以简化机器翻译知识获取难度，提高翻译精度和系统适应能力，减少人工编写规则和制作翻译词典的工作量，缩短研发周期，降低研制成本。

4.2 规则与统计相结合的机器翻译研究相关成果

基于规则和统计相结合的策略已经在诸如单词分割[26]、词性标注[27]、句法分析[28]、语块提取[29,30]、自动查错[31]、口语理解[32]、自动文摘[33]等研究领域取得成就。在机器翻译研究方面的成果也有很多，比如：文献[8]在1998年开发研制的英汉机器翻译系统BT863-2中，将英汉机器翻译中的歧义问题归纳为词法、兼类歧义、句法歧义和译文歧义等4种，研究了多层次渐进方式、规则与统计相结合的混合消歧策略。对词法和兼类歧义采用了基于规则的方法、在句法消歧中采用了GLR算法[35,36]和概率约束CFG语法相结合的策略、在译文消歧中使用基于目标语统计的词汇译文选择方法，达到了较好的效果。文献[37]在2002年提出了一种统计与规则相结合的目标语言生成策略，其在目标语言生成中，引入基于目标语言 N 元模型和句法关联关系信息约束机制的词项序位计算，增强了目标语言生成器对各词汇之间的内在关联关系的精确处理能力，减少了其对具体语言生成规则的依赖性，改善了译文的流畅度。文献[38]提出了两种结合统计方法和规则方法的优化策略，其一是运用SMT的解码器综合优化多引擎机器翻译结果，其二是运用浅层语言处理技术把SMT翻译模板进行加工，把处理结果当作RBMT子系统的词汇资源来利用，从而达到提高多引擎翻译系统的翻译精度的效果。

另一方面，规则和统计相结合的策略在SMT中也有很多应用。文献[30]采用统计和规则相结合的方法进行汉语的组块分析获得了很高的召回率。吴德恺提出的反向转换文法(ITG)和基于概率反向转换文法(SITG)的对位模型[39-41]，其实也是规则和统计相结合的产物之一。苏克毅还把统计手法在噪声信道模型中源语言到目标语言的转换归纳为9种形式[42]，分别为：词到词、短语到短语、语块到字符串、语块到语块、句法树到字符串、二叉树到二叉树、短语树到短语树、句法树到句法树、语义树到语义树。这种分类形式也显现了文法规则在SMT中的重要作用。

5 词义消歧中的理性主义和经验主义

词义消歧(Word Sense Disambiguation，WSD)的任务是确定一个多义词在给定的上下文语境中的具体含义，是自然语言理解的基础研究课题之一，在机器翻译、信息检索、文本处理、语音处理、语法和句法分析等领域应用广泛。自机器翻译研究诞生以来，词义消歧一直备受计算语言学家们的关注，它是自然语言处理中的一项艰巨任务，国内外有众多的学者致力于这项任务的研究工作。

词义消歧方法可分为基于知识的方法和基于统计的方法。基于知识的方法又可以分为基于规则的方法和基于词典的方法。基于统计的词义消歧方法按照机器学习方法可分为有监督(supervised)学习方法和无监督(unsupervised)学习方法。有监督词义消歧通常被看作词义分类问题，无监督词义消歧则被看作聚类问题。另外，由于语言学家提供的各类词典或知识库是获取词义消歧知识的重要来源，因此，人们往往也把基于词典或知识库的消歧方法单独区分开来进行研究。目前，基于词典的词义消歧方法是一种非常重要的消歧策略，如WordNet，Hownet，EDR电子词典等。近年来，运用bootstrapping，AdaBoost MH，Conditional Random Fields(CRF)，Co-Training等方法的半监督学习技术在这种方法中起到了重要作用。

典型的有监督学习方法有：基于互信息的消歧方法[43-49]、基于决策树模型的消歧方法[50]、基于贝叶斯分类器的消歧方法[51]、基于最大熵模型的消歧方法[52]和基于支持向量机[53](SVM)的消歧方法等。典型的无监督学习方法有：基于双语语料的词义消歧方法[54-64]、基于WEB的词义消歧方法[65]、基于聚类的词义消歧方法[66-75]等。其中，典型的基于聚类的词义消歧有：Latent Semantic Analysis(LSA)[67-69]，Hyperspace Analogue to Language(HAL)[70,71]，Clustering by Committee(CBC)[72]等等。典型的基于词典的分析方法有：基于机读词典的词义消歧、基于词典语义定义的方法[76]、基于义类词典的方法[77,78]、基于双语词典的方法[53,54]、基于领域信息的词义消歧方法[79,80]、基于百科知识(Wikipedia)的消歧方法[81]等等。其中，基于义类词典的方法又可以细分为：基于概念区域密度的词义分析方法[82-84]、基于结构化语义关系的图论式词义消歧方法[85-92]等等。文献[93,94]也对词义消歧做了综合性的论述。

由此可见，词义消歧方法也经历了基于规则的理性主义方法到基于统计的经验主义方法的变迁，并逐步走上融合规则、义类词典等知识库和多种统计方法相结合的集成消歧策略的道路。

另外，词义消歧的系统测评是词义消歧研究的重要环节之一，SENSEVAL是由国际计算语言学联合会(ACL)词汇兴趣小组(SIGLEX)于1997年开始组织的关于词义消歧的公共测评任务。第三次SENSEVAL评测结果显示，性能表现最好的系统在粗粒度定义下词义消歧的正确率和召回率依旧保持在79.3%，在细粒度定义情况下的正确率和召回率约为72.9%，性能表现排名前几位的系统分别采用了朴素贝叶斯分类器、SVM、最大熵方法，并且结合了多种知识库。从测试结果来看，词义消歧技术还有很大的改善空间[7]。目前，很多学者倾向于综合应用多种方法进行集成消歧，包括融合规则、义类词典等知识库、结合多种统计消歧方法，取得了良好的消歧效果。

结束语 综上所述，统计机器翻译的发展趋势是，统计机器翻译模型的改良需要融合更多的句法语义规则，以提高翻译质量。基于规则的机器翻译研究的发展趋势是，在既有翻译规则库达到某种特定程度时，结合各类复杂文法、非规范文法的具体特征，构建统计模型；建立特定文法与统计模型的相

关约束机制以实现两者的有机融合，优势互补，提高系统的可扩展能力和领域适应能力。两者在技术上的发展是殊途同归的。

本文介绍了自然语言处理技术中基于规则的理性主义方法和基于统计的经验主义方法的优缺点，以及机器翻译研究中基于规则的方法和基于统计的方法的优缺点；结合当前机器翻译研究的发展现状，提出了规则和统计相结合的机器翻译方法的基本框架，论述了词义消歧中的理性主义方法和经验主义方法的变迁和发展趋势。

机器翻译依旧面临众多难题，需要研究人员针对机器翻译研究任务中的具体问题进行分析、判断和取舍，拆分、整合和改进；需要研究人员不断进取，推动机器翻译研究的进步和发展。

参 考 文 献

[1] http://www.liktrans.com/www/news/2010-01/270.html

[2] http://news.sohu.com/20070102/n247395648.shtml

[3] 中国翻译产业迎来黄金发展期[N]. 人民日报海外版，2008-8-6

[4] Xu Jin-an. Prospects in Machine Translation[C]// 2010 Cross-Strait Conference on Information Science and Technology, CS-CIST 2010. Qinhuangdao, 2010: 368-372

[5] 冯志伟. 机器翻译研究[M]. 北京：中国对外翻译出版公司，2004

[6] 刘群. 汉英机器翻译若干关键技术研究[M]. 北京：清华大学出版社，2008

[7] 宗成庆. 统计自然语言处理 [M]. 北京：清华大学出版社，2008

[8] http://logos-os.dfki.de/

[9] John W, et al. An Introduction to Machine Translation[M]. Academic Press, 1992

[10] http://en.wikipedia.org/wiki/Weidner_Communications

[11] http://www.paho.org/ENGLISH/AM/GSP/TR/MACHINE_trans.htm

[12] White J S. Characteristics of the METAL Machine Translation System at Production Stage[C]// The 1st International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. Hamilton, 1985: 359-369

[13] Luckhardt H D. Sublanguages in machine translation[C]// Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics (Berlin). 1991: 306-308

[14] Arnold D. Eurotra: a European Perspective on MT [J]. Proceedings of the IEEE, Special Issue on Natural Language Processing, 1986, 74(7): 979-992

[15] Witkam T. DLT-an industrial R&D project for multilingual MT [C]// Proceedings of the 12th International Conference on Computational Linguistics. 1988: 756-759

[16] Uchida H. ATLAS II: A Machine Translation System Using Conceptual Structure as an Interlingua[C]// MachineTranslation Summit. Ohmsha, Tokyo, 1989: 93-100

[17] Nakamura J, Tsujii J, Nagao M. Grammar Writing System (GRADE) of Mu-Machine Translation Project and its Characteristics[C]// Proceedings of COLING 84. Stanford University, California, 1984: 338-343

[18] Nagao M. Present and future of machine translation systems[C]// Machine Translation Summit. Japan, September 1987: 84-86

[19] Satoshi S. MBT1: Example-based Word Selection[J]. Artificial Intelligence, 1991, 6(4): 592-600

[20] Satoshi S. MBT2: A Method for Combining Fragments of Examples in Example-Based Translation[J]. Artificial Intelligence, 1995, 75: 31-49

[21] Sumita E, Tsutsumi Y. A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching, TRI[R]. TR-87-1019. Tokyo Research Laboratory, IBM, 1988

[22] Brown R D. Example-based Machine Translation in the Pangloss System[C]// COLING-96: The 16th International Conference on Computational Linguistics. Copenhagen, 1996: 169-174

[23] http://www.mext.go.jp/a_menu/kagaku/chousei/1279442.html

[24] 米海涛，赵红梅，刘群. 第十二届机器翻译峰会和 NIST2009 机器翻译评测研讨会简介[J]. 中文信息学报，2009，23(6)：122-125

[25] http://www.hjtek.com/Article/ShowArticle.asp? ArticleID=33

[26] http://chasen.naist.jp/hiki/ChaSen/

[27] 周强. 规则与统计结合的汉语词类标注方法[J]. 中文信息学报，1995，9(2)：1-10

[28] 刘颖. 规则与统计结合分析汉语[J]. 计算机工程与应用，2002，38(7)：3-6

[29] 姜柄圭，张秦龙，谌贻荣，等. 面向机器辅助翻译的汉语语块自动抽取研究[J]. 中文信息学报，2007，21(1)：9-16

[30] 李素建，刘群，白硕. 统计和规则相结合的汉语组块分析[J]. 计算机研究与发展，2002，39(4)：385-391

[31] 张仰森，曹元大，俞士汶. 基于规则与统计相结合的中文文本自动查错模型与算法[J]. 中文信息学报，2006，20(4)：1-7

[32] 解国栋，宗成庆，徐波. 面向中间语义表示格式的汉语口语解析方法[J]. 中文信息学报，2003，17(1)：1-6

[33] 傅间莲，陈群秀. 基于规则和统计的中文自动文摘系统[J]. 中文信息学报，2006，20(5)：10-16

[34] 赵铁军，荀恩东，樊艳梅. 英汉机器翻译系统 BT863-Ⅱ的消歧策略研究[C]// 第五届中国人工智能联合学术会. CJCAI'98. 1998：207-212

[35] Tomita M. Efficient Parsing for Natural Language[M]. Norwell, MA, USA: Kluwer Academic Publishers, 1986

[36] Lavie A, Tomita M. GLR*-An Efficient Noise-Skipping Parsing Algorithm for Context-Free Grammars[M]// Bunt H, Tomita M, eds. Recent Advances in Parsing Technology, Text Speech and Language Technology series (vol. 1). Kluwer Academic Press, 1996

[37] 郭宏瞽，胡岗. 统计与规则相结合的目标语言生成策略[C]// 2002 年全国机器翻译研讨会. 2002：110-115

[38] Eisele A, Federmann C, et al. Hybrid machine translation architectures within and beyond the EuroMatrix project[C]// Proceedings of the 12th annual conference of the European Association for Machine Translation (EAMT 2008). Hamburg, Germany, September 2008: 27-34

[39] Wu D K. A Polynomial-Time Algorithm for Statistical Machine Translation[C]// Proceedings of ACL. 1996

[40] Wu D K. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora[J]. Computational Linguistics, 1997, 23(3): 377-403

[41] Wu D K, Wong H. Machine translation with a stochastic grammatical channel[C]// Proceedings of the ACL. 1998

[42] Su K Y. To Have Linguistic Tree Structures in Statistical Machine Translation? [C]// Proceedings of the 2005 12th IEEE International Conference on Natural Language Prcessing and Knowledge Engineering. 2005:3-6

[43] Brown P F, Della Pietra S A, Della Pietra V J, et al. Word-sense Disambiguation Using Statistical Methods[C]// Proceedings of the 29th ACL:264-270

[44] Brown P F, Della Pietra S A, Della Pietra V J, et al. Statistical Approach to Sense Disambiguation in Machine Translation[C]// Proceedings of the Fourth DARPA Workshop on Speech and Natural Language. Morgan Kaufman Publishers, 1991:146-151

[45] Carpuat M, Shen Y H, Yu X F, et al. Toward Integrating Word Sense and Entity Disambiguation into Statistical Machine Translation[C]// Proceeding of the International Workshop on Spoken Language Translation (IWSLT). 2006:37-44

[46] Carpuat M, Wu D K. Improving Statistical Machine Translation Using Word Sense Disambiguation[C]// Proceeding of 2007 Joint Conference on the EMNLP-CoNLL 2007. Prague, 2007: 61-72

[47] Carpuat M, Wu D K. Context-dependent Phrasal Translation Lexicons for Statistical Machine Translation[C]// Proceeding of Machine Translation Summit XI. Copenhagen, Sept. 2007:73-80

[48] Carpuat M, Wu D K. How Phrase Sense Disambiguation outperforms Word sense Disambiguation for Statistical Machine Translation[C]// Proceeding of the 11th Conference on Theoretical and Methodological issues in Machine Translation (TMI 2007). Sweden, Sept. 2007:43-52

[49] Chan Y S, Ng H T, Chiang D. Word Sense Disambiguation Improves Statistical Machine Translation[C]// Proceeding of the 45th Annual Meeting of the Association for Computational Linguistics (ACL). June 2007:33-40

[50] Preiss J. A detailed comparison of WSD systems: an analysis of the system answers for the SENSEVAL-2 English all words task[J]. Natural Language Engineering, 2006, 12(3):209-228

[51] Gale W A, Church K W, Yarowsky D. A Method for Disambiguating word senses in a large corpus[J]. Computers and Humanities, 1992, 26(5/6):415-439

[52] Chao G, Dyer M G. Maximum entropy models for word sense disambiguation[C]// Proceedings of 19th International Conference on Computational Linguistics (COLING-02). 2002:1-7

[53] Dagan I, Itai A, Markovitch S. Two languages are more informative than one[C]// Proceeding of the 29th ACL. 1991:130-137

[54] Dagan I, Itai A. Word sense disambiguation using a second language monolingual corpus[J]. Computational Linguistics, 1994, 20(4):563-596

[55] Resnik P, Yarowsky D. A perspective on word sense disambiguation methods and their evaluation[C]// Proceeding of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How. Morristown: Association for Computational Linguistics, 1997:79-86

[56] Eseudcro G, Marquez L, Rigau G. Boosting applied to word sense disambiguation[C]// Proceeding of the 12th European Conference on Machine Learning. Berlin, HeideIberg: Springer-Verlag, 2000:129-141

[57] Ide N, Ejavec T, Tufts D. Sense discrimination with parallel corpora[C]// Proceeding of the ACL SIGLEX Workshop on word Sense Disambiguation: Recent Successes and Future Directions. Morristown: Association for Computational Linguistics, 2002: 54-60

[58] Li C, Li H. Word translation disambiguation using bilingual bootstrapping[C]// Proceeding of the 40th ACL. 2002:343-351

[59] Ng H T, Wang B, Chan Y S. Exploiting parallel texts for word sense disambiguation: An empirical study[C]// Proceeding of the 41st ACL. 2003:455-462

[60] Diab M, Reanik P. An unsupervised method for word sense tagging using parallel corpora[C]// Proceeding of the 40th ACL. 2002:255-262

[61] Diab M. Word sense disambiguation within a multilingual framework[D]. University of Maryland College, 2003

[62] Diab M. An unsupervised approach for bootstrapping Arabic word sense tagging[C]// Proceeding of the Arabic Based Script Languages. COLING 2004

[63] Diab M. Relieving the data acquisition bottleneck in word sense disambiguation[C]// Proceeding of the 42th ACL. 2004:303-310

[64] Bhattacharya I, Getoor L, Bengio Y. Unsupervised sense disambiguation using bilingual probabilistic' models[C]// Proceeding of the 42nd ACL. 2004:287-294

[65] Klapaftis I P, Manandhar S. Google & WordNet based word sense disambiguation[C]// Proceeding of the 22nd ICML Workshop on Learning & Extending Ontologies. New York: Association for Computing Machinery, 2005

[66] Schutze H. Automatic word sense discrimination[J]. Computational Lingnistics, 1998, 24(1):97-123

[67] Deerwester S, Dumais S T, Fumas G W, et al. Indexing by latent semantic analysis[J]. Journal of the Amedean Society for Information Science, 1990, 41(6):391-407

[68] Landauer T K, Dumais S T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge[J]. Psychological Rfview, 1997, 104 (2):211-240

[69] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis[J]. Discourse Processes, 1998, 25(2):259-284

[70] Burgess C, Lund K. Modefing parsing constraints with high-dimensional context space[J]. Language and Cognitive Processes, 1997, 12(2/3):177-210

[71] Burgess C, Lurid K. The dynamics of meaning in memoryL[C]// Dietrich E, Markman A, eds. Cognitive Dynamics: Conceptual Representational Change in Humans and Machines. 2000: 117-156

[72] Lin D K, Pantel P. Concept discovery from text[C]// Proceeding of the 19th International Confference on Computational Linguistics(COLING-2002). 2002:577-583

[73] Pedersen T, Bruce R. Distinguishing word senses in untagged text[C]// Proceeding of the 2nd Conference on Empirical Methods in Natural Language Processing. 1997:197-207

[74] Pedersen T. Knowledge lean word sense disambiguation[C]// Proceeding of the 15th National Conference on Artificial Intelligence. 1998:800-805

[75] Purandare A, Pedersen T. Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces[C]// Proceedings of the Conference on Computational Natural Language Learning (CoNLL). Boston, MA: Quinlan Publishing, May 2004:41-48

(下转第 250 页)

Founctions[J]. Artificial Intelligence,2006,170 (11):909-924
[17] 涂嘉文,徐守时. 贝叶斯方法与 Dempster-Shafer 证据理论的讨论[J]. 红外与激光工程,2001,30(2):61-64
[18] 王晓丽. 一种混合结构的数据融合算法研究及在目标识别中的应用[D]. 秦皇岛:燕山大学,2004
[19] 徐从富,耿卫东,潘云鹤. 面向数据融合的 DS 方法综述[J]. 电子学报,2001,29(3):393-396
[20] 徐从富,耿卫东,潘云鹤. Dempster-Shafer 证据理论方法理论与应用的综述[J]. 模式识别与人工智能,1999,12(4):424-430
[21] 刘同明. 基于证据理论模糊推理的多传感器信息融合海上目标识别[J]. 模识别与人工智能,1999,12(1):25 -31
[22] Hansen L K, Salamon P. Neural network ensembles[J]. IEEE Trans on Pattern analysis and machine intelligence, 1990, 12 (10):993-1001
[23] Schapire R E. The strength of weak learnabiliy[J]. Machine Learning,1990,5(2):197-227
[24] Sewell M. Ensemble learning [R]. University College Lodon, 2007
[25] Hampshire J, Waibel A. A novel objective function for improved phoneme recognition using time-delay neural networks[J]. IEEE Trans Neural Networks,1990,1(2):216-228
[26] Schapire R E, Singer Y, Singhal A. Boosting and Rocchio appliedto text filtering[C]// Proc the 21st Annual ACM SIGIR International Conference on Research and Development inInformation Retrieval. NY,1998:215-223
[27] Tumer K, Ghosh J. Error correlation and error reduction in ensemble classifiers[J]. Connection Science,1996,8(3/4):385-403
[28] Dietterich G, learning E. The Handbook of Brain Theory and Neural Networks(Second Edition)[M]. MIT press,2002:385-577
[29] 刘明,袁保宗,等. 从局部分类精度到分类置信的变换[J]. 计算机研究与发展,2008,45(9):1612-1619

(上接第 229 页)

[76] Lesk M E. Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone[C]// Proceedings of the ACM SIGDOC Conference. Toronto, Ontario,1986:24-26
[77] Walker D E. Knowledge Resource Tools for Accessing Large Text Files[C]// Proceedings of Conference on Theoretical and Methodological Issues in Machine Translation of Natural Laguages. Colgate University, Hamilton, New York, August 1985:335-347
[78] Yarowsky D. Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora[C]// Proceedings, COLING-92. Nantes, France,1992:241-246
[79] Dewey M, et al. Dewey Decimal Classification and Relative Index (22 edition)[M]. OCLC Online Computer Library Center, October 2003
[80] Gonzalo J, Verdejo F, Peters C, et al. Applying eurowordnet to cross-language text retrieval[J]. Eurowordnet: A Multilingual Database With Lexical Semantic Networks,1998,32:185-207
[81] Mihalcea R. Using Wikipedia for automatic word sense disambiguation[C]// Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, New York, April 2007
[82] Agirre E, Rigau G. A proposal for Word Sense Disambiguation Using Conceptual Distance[C]// Proceedings of the First International Conference on Recent Advanced in NLP. Bulgaria, 1995:162-171
[83] Rosso P, Masulli F, Buscaldi D, et al. Automatic noun sense disambiguation[C]// Proceeding of CICLing. 2003:273-276
[84] Buscaldi D, Rosso P, Masulli F. Integrating Conceptual Density with Word-Net Domains and CALD Glosses for Noun Sense Disambiguation[J]. Lecture Notes in Artificial Intelligence, 2004,3230:183-194
[85] Litkowski K. Use of machine readable dictionaries for word-sense disambiguation in senseval-2 [C] // Proceedings of the Senseval-2 Workshop. Toulouse,2001:107-110
[86] McCarthy, Diana, Koeling R, et al. Using automatically acquired predominant senses for word sense disambiguation[C] // Proceedings of the ACL Senseval-3 Workshop. Barcelona, Spain, 2004:151-154
[87] Morris J, Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text[J]. Computational Linguistics,1991,17(1):21-48
[88] Galley M, McKeown K. Improving word sense disambiguation in lexical chaining[C]// Proceeding of the 18th International Joint Conference, on Artificial Intelligence, UCAI 2003. 2003: 1486-1488
[89] Mihalcea R, Tarau P, Figa E. PageRank on semantic networks with application to word sense disambiguation[C]// Proceeding of the 20th International Conference on Computational Linguistics (COLING-2004)
[90] Navigli R, Velardi P. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(7):1075-1086
[91] Agirre E, Martlnez D, de Lacalle O L, et al. Two graph-based algorithms for state-of-the-art WSD[C]// Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Barcelona: Association for Computational Linguistics,2006:583-593
[92] Sinha R, Mihalcea R. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity[C]// Proceeding of the IEEE International Conference on Semantic Computing (ICSC). Washington: IEEE Computer Society,2007: 363-369
[93] Navigli R. Word sense disambiguation: A survey[J]. ACM Computer. Survey,2009,41(2)
[94] Wu Yun-fang. A survey of Chinese word sense disambiguation: Resources, methods and evaluation[J]. Journal of Contemporary Linguistics,2009,11(2):113-123
[95] 刘群. 统计机器翻译综述[J]. 中文信息学报,2003,17(4):1-12
[96] 熊德意,刘群,林守勋. 基于句法的统计机器翻译综述[J]. 中文信息学报,2008,22(2):28-39