

Prospects in Machine Translation

Jin'an Xu

School of Computer and Information Technology, Beijing Jiaotong University, Beijing

Email: xja2010@gmail.com

Abstract: This paper firstly introduces three main methods of constructing machine translation system, and describes recent development of machine translation technologies, and discusses prospects of future development. Meanwhile, some latest machine translation systems will be mentioned.

Keywords: machine translation; nature language processing; computational linguistics

机器翻译展望

徐金安

北京交通大学计算机与信息技术学院, 北京, 100044

Email: xja2010@gmail.com

摘要: 本文介绍了三种主流机器翻译方法, 结合机器翻译的实现方法和当前的机器翻译产品市场形势, 描述了机器翻译技术的发展现状, 并进行了总结和展望。

关键词: 机器翻译; 自然语言处理; 计算语言学

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, 对翻译的潜在需求十分巨大。据美国 ABI 公司对世界翻译市场的调查结果, 2005 年的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经达到了 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。

本文结构如下: 在第二部分介绍三种主流机器翻译方法; 第三部分分析了机器翻译现状; 第四部分探讨规则和统计方法相结合的研究策略; 最后, 对机器翻译研究进行总结和展望。

2 机器翻译方法

2.1 机器翻译发展

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, 是目前国际上最具有挑战性的研究课题之一。

按照冯志伟老师的说法^[4], 到 19 世纪 80 年代, 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等。近年来, 在 ACL 等主流学术刊物上, 多数成果都集中在基于大规模语料库的统计机器翻译方法上, 令人瞩目。

目前, 基于规则的方法、基于实例的方法和基于统计的方法占据机器翻译的主流地位。多引擎机器翻译策略促进了三种主流机器翻译系统的平衡发展。

资助信息: 中央高校基本科研业务费专项资金资助

2.2 基于规则的机器翻译 (Rule-Based Machine Translation, RBMT)

RBMT 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。

RBMT 方法对句子结构以及长距离依存关系的处理能力很强, 尤其对语言现象明确和句法结构相对规范的源语言句子, 显现出很强的处理能力和很好的翻译结果。此外, 还具有对知识表达的抽象程度高、代表性强、对不同语料的覆盖率高、系统运行占用资源少等优点。其主要缺点是: 语法的分析和生成规则主要由人工编写, 知识获取难、工作量大、研制成本高, 规则的主观性强, 规则的一致性难以保障, 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^[5]。

2.3 基于实例的机器翻译 (Example-Based Machine Translation, EBMT)

EBMT 方法的基本思想由日本著名机器翻译专家长尾真提出, 其基本思想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]。

EBMT 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: 第一, 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; 第二, 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, 导致泛化和匹配的平衡产生问题; 第三, 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]。

2.4 基于统计的机器翻译 (Statistical Machine Translation, SMT)

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 等人在提出了 5 种复杂程度递增的数学模型。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。其后, 吴德凯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。Yamada 和 Knight 等在 IBM 的统计翻译模型的基础上, 提出了一种基于句法结构的统计翻译模型。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, 该模型具有灵活性强、可以随意增加特征、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。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, 典型的翻译模型有串到树、树到树、树到串翻译模板等等。

SMT 方法具有良好的数学模型、无指导的学习能力和良好的鲁棒性; 获取知识的人工成本低、不需要依赖大量的语言知识, 直接依靠统计结果进行语义消歧处理和译文选择, 从而避开了语言理解的诸多难题, 同时大大缩短了系统的研制周期。其主要问题在于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。翻译模型和语言参数的精确性直接依赖于语料的多少, 而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。

2.5 多引擎机器翻译 (Multi-engine Machine Translation)

1994 年, Frederking 提出了一种多翻译引擎的方法, 1998 年 Hogan 通过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。缺点是系

统规模庞大、造价高、不易维护。

3 国内外机器翻译现状

在机器翻译的实用化方面,国内外相关产品的特点是:翻译语言和服务形式多样化、系统大规模集成化、使用终端小巧化和自动语音翻译系统实用化。

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、跨国公司等等。

基于规则的翻译系统还有很多,比如法国纺织研究所曾经研制的 Titus4、美国的 Weidner 和 Paho、德国的 Metal 和 Susy、欧盟的 Eurotra、荷兰 BSO 公司的 DLT、日本的 Atlas 和 Mu 等等。

早期的 EBMT 系统有:日本的 MBT1 和 MBT2、日本 ATR 的 ETOC 和 EBMT 系统、美国的 Pangloss 系统中的 EBMT 子系统等等。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 方法。

目前, SMT 系统产品为数不多,较知名的有 Google 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,统计模型不断地得到改善,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)、德国西门子公司等机构。最近,便携式自动语音翻译系统产品相继问世。该类产品有以下特点: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;现在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 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^[7]。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,机器翻译先后经历了 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,进入商品化进程,90 年代开始进入网络化时期,21 世纪初开始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。

4 理性主义方法与经验主义方法的融合

机器翻译研究的发展历程凸显了自然语言处理的方法论之争,促使基于规则的理性主义方法和基于统计的经验主义方法从对立走向结合之路。1990 年以前,基于规则的理性主义方法占据主导地位,发展很快,且日趋成熟。但是人们也发现 RBMT 系统的性能很难进一步提高。1990 年之后,基于统计的经验主义机器翻译方法得到了强势发展。目前, RBMT 研究注

重如何借鉴统计方法从大规模语料库中自动获取翻译知识和翻译规则;而 SMT 研究注重融合更多的语言结构和语法知识,以提高翻译质量。

规则方法可精确地描述语言特征,而统计方法能实现翻译知识的自动获取。各种方法对译文质量的保障程度深浅不一,翻译精度也还不尽人意,凸现了理性主义方法和经验主义方法之间的矛盾和对立,也蕴涵着在方法论层面上仍存在较大的发展空间,两者的有机结合成为机器翻译研究人员普遍关注的热点。

早期的 RBMT 系统的核心问题是如何构造比较完备而且适应性较强的规则库。早期的规则大都采用非此即彼的确定性规则,而规则的增加导致彼此冲突;翻译规则本身很难建立系统化的分类标准,难以找到恰当的规则粒度来描述抽象的语言特征,导致系统的领域适应能力差,适用于新领域时往往要求整套翻译规则重写,代价和成本昂贵。

RBMT 系统在基于大规模语料库的机器翻译知识自动获取研究方面,越来越多地借鉴统计自然语言处理技术。比如,聚类算法、相关算法、复杂特征集和合一运算、概率上下文无关文法、N 元模型、HMM、朴素的贝叶斯方法、SVM、决策树模型、最大熵模型,基于错误驱动的转换方法、神经网络等等。最近,运用 bootstrapping、Conditional Random Fields(CRF)和 Co-Training 等方法,在小规模标注语料库的指导下,基于半监督机器学习解决问题的研究也有所成就。

基于规则和统计相结合的策略已经在诸如单词分割、词性标注、句法分析、语块提取、自动查错、口语理解、自动文摘、机器翻译等方面都取得了研究成果。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^[9]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,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^[10]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^[11]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^[12-14],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^[15],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。

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。RBMT 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实现两者的有机融合,进行优势互补,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。

5 总结和展望

本文介绍、分析和评价了三种主流机器翻译方法;同时,阐述了国内外机器翻译发展现状和最新动态;从方法论和技术层面探讨了机器翻译面临的问题和发展趋势。

机器翻译依旧面临众多难题,任重而道远,充满挑战性,值得我们锲而不舍地上下求索、刻苦钻研。

References (参考文献)

- [1] <http://www.liktrans.com/www/news/2010-01/270.html>
- [2] <http://news.sohu.com/20070102/n247395648.shtml>
- [3] Chinese translation industry usher in golden period of development, the People's Daily Overseas Edition, August 6, 2008.
- [4] Zh. Feng, Research on Machine Translation. Beijing: China Translation and Publishing Company, 2004.
- [5] Ch. Zong, Statistical Natural Language Processing, Tsinghua University Press, May 2008 first edition.
- [6] Q. Liu Several Key Technologies of CE Machine Translation,

- Tsinghua University Press, October 2008 first edition.
- [7] <http://www.hjtek.com/Article/ShowArticle.asp?ArticleID=33>
 - [8] X. Liu, WSD Methods of EC Machine Translation, Ph.D. thesis, Harbin Institute of Technology, Jan. 1998.
 - [9] H. Guo, G. Hu, Combining Statistical Model with Linguistic Knowledge in Target Language Generation. The 4th China Workshop on Machine Translation, 2002, pp110-115.
 - [10] A. Eisele, C. Federmann, H. Uszkoreit, H. Saint-Amand, M. Kay, M. Jellinghaus, S. Hunsicker, T. Hermann, and Y. Chen., Hybrid machine translation architectures within and beyond the EuroMatrix project. In Proceedings of the 12th annual conference of the European Association for Machine Translation (EAMT 2008), pages 27-34, Hamburg, Germany, September 2008.
 - [11] J. Li, Q. Liu, Sh. Bai, Chinese Chunking Parsing Using Rule-Based And Statistics-Based Methods, Journal of Computer Research And Development, 2002 39(4), pp.385-391.
 - [12] Wu D. K., A Polynomial-Time Algorithm for Statistical Machine Translation. In Proceedings of ACL 1996.
 - [13] Wu D. K., Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics, 23(3):377-403, 1997.
 - [14] Wu D. K. and Wong H., Machine translation with a stochastic grammatical channel. In Proceedings of the ACL, 1998