
2018

机器翻译 与人工智能研究报告

AMiner 研究报告第五期

清华大学计算机系-中国工程科技知识中心
知识智能联合研究中心 (K&I)

2018 年 5 月

Contents 目录

1 概念篇

1.1 机器翻译简介	2
1.2 机器翻译发展历程	2
1.3 我国机器翻译现状	5

2 技术篇

2.1 理性主义方法	9
2.1.1 基于规则的机器翻译	9
2.2 经验主义方法	12
2.2.1 基于统计的机器翻译	13
2.2.2 基于实例的机器翻译	15
2.2.3 基于深度学习的机器翻译	16

3 人才篇

3.1 领军人物	22
3.2 中坚力量	28
3.3 领域新星	33

4 应用篇

4.1 趋势篇	36
---------------	----

5 趋势篇

5.1 趋势篇	39
---------------	----

图表目录

图 1 抽象转换的分层实现	3
图 2 机器翻译技术源头	4
图 3 机器翻译过程	9
图 4 机器翻译的转换层面	10
图 5 直接翻译过程	10
图 6 基于转换方法的翻译流程	11
图 7 中间语言与转换方法比较	12
图 8 中间语转换翻译过程	12
图 9 统计机器翻译典型模型	13
图 10 基于统计的机器翻译模型	14
图 11 基于实例方法翻译过程	15
图 12 深度学习发展脉络	16
图 13 机器翻译领域全球学者分布	21
图 14 机器翻译领域中国学者分布图	21
图 15 机器翻译各国人才顺逆差图	22
图 16 机器翻译领域全局热度	39
图 17 机器翻译领域近期热度	39

摘要

随着计算机科学技术的发展,机器翻译作为自然语言处理研究的重要组成部分越发受到人们关注。经过了几十年的努力,以机器翻译为代表的自然语言处理工作取得了巨大的进展,并且在未来有着广阔的发展空间,为了梳理机器翻译领域的研究概括,我们编写了此份报告,主要包括:

机器翻译概论。首先对机器翻译进行定义,接着对机器翻译的发展历程进行了梳理,对我国机器翻译现状进行了简单介绍。

机器翻译技术原理。机器翻译的技术原理可以概括为基于理性主义的方法和基于经验主义的方法两种,分别对两种方法下的基于规则的翻译方法、基于实例的翻译方法、基于统计的翻译方法以及基于深度学习的翻译方法进行介绍。

机器翻译领域专家介绍。利用 AMiner 大数据对机器翻译领域专家进行深入挖掘,选取国内外有代表性的专家进行简要介绍。

机器翻译的应用及趋势预测。机器翻译在现实生活中应用广泛,在文本翻译、语音翻译、图像翻译和视频、VR 翻译等领域均有了不同的进展,在此基础上,对机器翻译未来的发展趋势做出了相应的预测。

1 concept

概念篇



1 概念篇

1.1 机器翻译简介

机器翻译（Machine Translation）是指运用机器，通过特定的计算机程序将一种书写形式或声音形式的自然语言，翻译成另一种书写形式或声音形式的自然语言。机器翻译是一门交叉学科（边缘学科），组成它的三门子学科分别是计算机语言学、人工智能和数理逻辑，各自建立在语言学、计算机科学和数学的基础之上。

机器翻译可以实现世界上不同国家不同语言间的低成本交流，其主要优点体现为：

- **成本低。**相对于人工翻译来说，机器翻译的成本要低很多。机器翻译需要人工参与的程序其实很少，基本上由计算机自动完成翻译，大大降低了翻译成本。
- **易把控。**机器翻译的流程简单快捷，在翻译时间的把控上也能进行较为精准的估算。
- **速度快。**计算机程序的运行速度非常快，其速度是人工翻译速度不可比拟的。

由于这些优点，机器翻译在这几十年来得到了快速的发展。在具体应用上一般分为三种，分别是：词典翻译软件、计算机辅助翻译软件和机器翻译软件。

第一种是最基本的网络查词翻译，查询对象一般为单个的字词、简单的词组或者是固定结构。

第二种为计算机辅助翻译，英文简称 CAT（Computer Aided Translation），其原理为利用计算机的记忆功能将译者之前翻译的资料进行整理，以便为之后出现的类似翻译提供便利条件。CAT 软件产业已经比较成熟，例如 Google Translator Toolkit、Microsoft LocStudio 等，Trados（塔多思）占有国际计算机辅助翻译软件产业绝大多数的市场份额，微软、西门子等国际大公司都是它的用户。

第三种是机器翻译软件，也叫做计算机翻译，即 MT（Machine Translation）。其原理为应用计算机按照一定规则把一种自然语言转换为另一种目标自然语言。此过程一般指自然语言之间句子和段落等的翻译，大部分见诸于世的翻译软件，如谷歌翻译、金山词霸和有道翻译等均属于机器翻译软件。

1.2 机器翻译发展历程

机器翻译思想正式提出于 1949 年，Warren Weaver 发表《翻译》备忘录，在那以后至今的时间里，机器翻译研究经历了一个曲折的发展过程。

第一台数字电子计算机诞生于 1946 年，从那以后，人们就开始思索如何运用计算机代替人从事翻译工作的问题，甚至在此之前，图灵就已经开始思考计算机是否能够进行思维这一问题。1949 年，信息论先驱 Warren Weaver 发表了有关机器翻译的备忘录，提出了机器翻译的可计算性，他提出两个主要观点：第一，他认为翻译类似于解读密码的过程；第二，他认为原文与译文“说的是同样的事情”，因此，当把语言 A 翻译为语言 B 时，就意味着从语言 A 出发，经过某一“通用语言”或“中间语言”，可以假定是全人类共通的。这是机器翻译发展初始阶段的第一件标志性事件；1954 年美国乔治敦大学（Georgetown）在 IBM 的协同下进行的英俄翻译实验开始了，在翻译自动化方面的尝试是机器翻译发展初始阶段的第二

件标志性事件。

总体来说，这一阶段人们头脑中已经形成了机器翻译的概念，并且已经意识到可以利用语法规则的转换和字典来实现翻译目的。人们乐观地认为只要扩大词汇量和语法规则，在不久的将来，机器翻译问题会比较完美地得以解决。所以在此之后的很长一段时间，全球各国大力支持机器翻译项目，一个机器翻译研究的高潮就此形成。

好景不长，1966 年 11 月，美国语言自动处理咨询委员会（ALPAC）从机器翻译的速度、质量、花费以及当时人们对机器翻译的需求等几个角度，对当时的各个翻译系统进行了一次评估，公布了著名的 APLAC 报告，给机器翻译研究工作浇了一盆凉水。报告提出，机器翻译的译文质量明显远低于人工翻译，难以克服的“语义障碍”是当时机器翻译遇到的问题，这份报告全面否定了机器翻译的可行性，建议各大机构停止对机器翻译的投资和研究。尽管报告的结论过于仓促、武断，但是这一阶段关于机器翻译的研究的确没有解决许多至关重要的问题，并没有对语言进行深入的分析。此后在世界范围内，机器翻译出现了空前的萧条局面。

20 世纪 80 年代末，由于微处理器的出现，计算机能力获得了突飞猛进的发展，机器翻译这一学科有着极大的开发潜力和经济利益，被重新提起。许多大公司开始投入资金和人力进行研究，使得机器翻译得到了复苏和重新发展的机会。这一时期，计算机和语言学的一些基础工作，比如许多重要的算法的研究已经到达了一个比较深入的阶段，对语法和语义的研究也已经有了一些比较重大的成果，词法分析、句法分析的算法相继得到开发，并且加强了软件资源，例如电子词典的建设。翻译方法以转换方法为代表，开始普遍采用以分析为主，辅以语义分析的基于规则方法来进行翻译，采用抽象转换表示的分层实现策略，如图 1。语法与算法的分开是这一时期机器翻译的另一个特点。所谓语法与算法分开，就是指把语言分析和程序设计分开来成为两部分操作，程序设计工作者提出规则描述的方法，而语言学工作者使用这种方法来描述语言的规则。

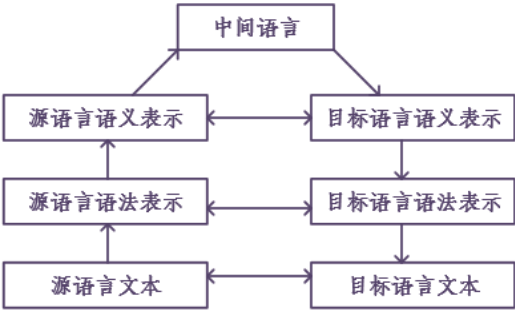


图 1 抽象转换的分层实现

现在，机器翻译已经成为世界自然语言处理研究的热门。原因之一是网络化和国际化对翻译的需求日益增大，翻译软件商业化的趋势也非常明显。这一时期的翻译方法我们一般称之为基于经验主义的翻译方法，主要是基于实例和基于统计的方法，特点是注重大规模语料库的建设，开始了针对大规模的真实文本处理。同时，这一阶段的研究工作开始解决一个比文本翻译更加复杂和艰难的问题——语音翻译。而且由于 Internet 上的机器翻译系统具有巨大的潜在市场和商业利益，此时网上翻译机器系统也进入了实用领域的新突破阶段。

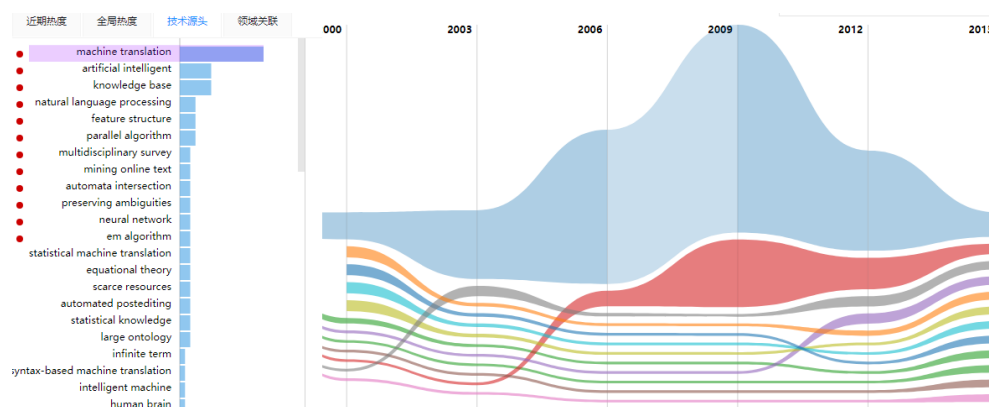


图 2 机器翻译技术源头

机器翻译功能越来越强大，从最初只能进行简单的单词翻译，到之后可以翻译出基本符合语法的句子，慢慢可以翻译具有一定逻辑性的句子，现在部分软件已经可以自主联系上下文进行翻译，翻译结果的准确性与可读性都已经取得了非常大的进步。

近年来，加入了“深度学习技术”等人工智能的机器翻译已经不止于简单的将一个个单词翻译成另一种语言，而是可以像人工翻译一样，不断向前回顾理解结构复杂的句子，同时联系上下文进行翻译。最为明显的就是现在的部分机器翻译软件已经可以理解每一个代词具体指代谁，这在许多年前是不可想象的。

实现这种功能的关键，分别依赖于两种神经网络架构，一个是循环神经网络（RNN，Recurrent Neural Networks），另一个是卷积神经网络（CNN，Convolutional Neural Network），目前关于两种网路架构哪种更适用于机器翻译的争论还有很多。

● 循环神经网络

循环神经网络的关键在于“循环”二字，计算机系统会“记住”上一次输出的内容，并以此来决定下一次输出。有了上一次和下一次的概念，神经网络就不会把输入和输出的信息看作是独立的，而是相互关联的时间序列。这样可以通过以往的序列关联猜测到下一个序列会出现的词。在翻译时，神经循环网络把源语言当作输入序列，把翻译语言当作输出序列，由于每次的输出都会参考上一次输出的结果，所以机器翻译更具有整体性，可读性和准确性更高，而不是简单地翻译单词。目前，循环神经网络运用最为熟练的应该是谷歌翻译，谷歌曾提出利用神经网络系统进行机器翻译，据称汉译英的错误率最高下降了 85%。

● 卷积神经网络

卷积神经网络可以同时处理多个语言片段，并且具有信息分层处理能力。将文本序列化、单词向量化，经过分层处理后再输出结果。在分层过程中，还会不断回顾原文本来确定下一个输出序列。提出这种技术的是 Facebook 和最近的机器翻译新秀 DeepL。2017 年上半年，Facebook 宣布推出了基于卷积神经网络开发的语言翻译模型，据说比基于循环神经网络开发的语言翻译模型速度可以快 9 倍，而且准确率更高。在测试上，Facebook 翻译系统在英语-德语、英语-法语的测试上都比循环神经网络更接近人工翻译。

不管是哪种系统，都不是机器翻译的终点，比如谷歌近期提到的不再基于卷积神经网络的注意力机制，以及多层神经网络、深度神经网络等，都是解决机器翻译问题的探索，在速

度、计算机资源消耗、情感理解等多种维度上各有不同的表现。

1.3 我国机器翻译现状

中国的机器翻译研究始于 20 世纪 50 年代，但是由于国际环境和电脑发展水平的束缚，国内真正对机器翻译的研究是在 20 世纪 80 年代晚期。具有重要意义的标志性成果是著名的“863 智能英-汉翻译系统”。20 世纪 90 年代，随着电脑技术的发展和对外交流的扩大，机器翻译的使用变得日趋频繁；机器翻译不仅是必要的，而且随着机器翻译软件发展到了前所未有的新高度，机器翻译也成为可能。机器翻译研究形成了独立研究机构和政府研究组织共存的良好面貌。国内成功的机器翻译的开发也呈现出前所未有的繁荣景象。

第一阶段的开发期是在 1957 年，中科院语言研究所、电脑科技研究所与中俄机器翻译合作，成功译出了九类复杂的句式。作为世界上的两种重要语言，英汉互译是国内外诸多学者所关心的。

第二阶段由于政治原因和机器翻译固有的困难而停滞。在此阶段，汉英机器翻译研究几乎止步不前。

第三阶段是大发展阶段，始于 1975 年。国内的机器翻译列入了“六五”“七五”“863”等主要研究计划。研究者集中精力进行了多个科研院所的协作研究，开展了与国际研究机构的合作和沟通，不仅培养了大批人才，积累了资源，而且把我国机器翻译带入了繁荣期。

上世纪 90 年代以来，我国相继推出了一系列机器翻译软件，例如“译星”“通译”等。随着市场需求的扩大，机器翻译成为一种新兴产业，走向了专业化和市场化。

近几年国内机器翻译发展很快，各大 IT 公司都相继推出自己的机器翻译系统，而且神经翻译技术和深度学习技术作为一种新的机器翻译范式，在诸多语种及应用场景中的翻译质量已经超越了统计机器翻译技术，并成为目前学术界和工业界研究的热点，以下对各大 IT 公司机器翻译进展逐一介绍。

2010 年初，百度组建了机器翻译核心研发团队，2011 年 6 月 30 日，百度机器翻译服务正式上线，目前，百度翻译支持全球 28 种语言互译、756 个翻译方向，每日响应过亿次的翻译请求。此外，百度翻译还开放了 API 接口，目前已有超过 2 万个第三方应用接入。华为、OPPO、中兴、三星等手机厂商，金山词霸、灵格斯词霸、敦煌网等众多产品均接入了百度翻译 API。百度还将基于神经网络的机器翻译引入机器翻译中，这一应用比谷歌翻译要早一年，在海量翻译知识获取、翻译模型、多语种翻译技术等方面取得重大突破，实时准确地响应互联网海量、复杂的翻译请求。其所研发的深度学习与多种主流翻译模型相融合的在线翻译系统以及基于“枢轴语言”等技术，处于业内领先水平，在国际上获得了广泛认可。

科大讯飞成立之时就再布局语言和翻译领域布局项目。基于深度神经网络算法上的创新和突破，科大讯飞在 2014 年国际口语翻译大赛 IWSLT 上获得中英和英中两个翻译方向的全球第一名；在 2015 年又在由美国国家标准技术研究院组织的机器翻译大赛中取得全球第一的成绩。2017 年科大讯飞还推出了多款硬件翻译产品，其中晓译翻译机 1.0plus 将世界上最先进的神经网络翻译系统，从在线系统优化成一个离线系统。它可以在没有网络的情况下提

供基本的翻译服务。

阿里巴巴 2015 年收购了国内最大的众包翻译平台——365 翻译，开始涉入机器翻译领域。2016 年 10 月起正式开始自主研发 NMT 模型，2016 年 11 月首次将 NMT 系统的输出结果应用在中英消息通讯场景下的外部测评中，并取得了不错的成绩。2017 年初阿里正式上线了自主开发的神经网络翻译系统，为阿里经济体复杂多样的国际化需求提供可靠的技术支撑。阿里机器翻译是基于阿里巴巴海量电商数据并结合机器学习、自然语言处理技术，实现多语言语种识别与自动翻译功能，为跨境电商信息本地化与跨语言沟通上提供精准、快捷、可靠的在线翻译服务，其宗旨是“让商业没有语言障碍”。

2016 年初，腾讯开始研发 AI 翻译产品，并正式推出机器人翻译——翻译君，支持中英日韩法德意土等 15 种语言和 80 个语种的对翻译。2017 年宣布翻译君上线“同声传译”新功能，用户边说边翻的需求得到满足，语音识别+NMT 等技术的应用保证了边说边翻的速度与精准性。腾讯机器翻译基于腾讯领先的底层算法、丰富的中文知识图谱和先进的 NLP 引擎能力，结合了神经网络机器翻译和统计机器翻译的优点，对源语言文本进行深入理解，使翻译效果更为准确，同时支持语音翻译、图片翻译、语种识别等多种场景，大大减轻传统文本翻译的读写成本。

机器翻译是搜狗人工智能战略中的重要一环，一方面可以满足用户在搜索过程中大量机器翻译需求，一方面还可以通过搜索和机器翻译技术的结合，帮助中文用户打破语言障碍，搜索并浏览全世界外语信息。2016 年 5 月 19 日，搜狗正式上线英文搜索。搜狗英文搜索提供跨语言检索功能，可自动将中文翻译成英文进行查询，再生成英文查询结果。对于不擅长英文的用户，可以节省很多“先翻后搜”的搜索时间。2017 年 11 月的乌镇世界互联网大会上，搜狗展示了机器同传技术，可将演讲者的中文同步翻译成英文并实时上屏。12 月 21 日，搜狗英文搜索正式升级为搜狗海外搜索频道，并同步上线了搜狗翻译频道。2018 年 3 月，搜狗上线定位旅游用的翻译机——翻译宝，开始了在机器翻译领域硬件的探索。目前，搜狗已经上线了基于神经网络的机器翻译频道，并发布了跨语言搜索系统，为用户提供高质量的英文网页搜索服务，并同时能够将搜索结果翻译为中文帮助用户理解。

网易 2011 年创立网易感知与智能中心，拥有自建分布式深度学习平台，其自主研发的图像处理、语音识别、智能问答等 AI 技术，已经在有道翻译中得到了应用和推广。2017 年 5 月网易有道在 GMIC 未来创新峰会上公布：由网易公司自主研发的神经网络翻译技术正式上线。此次在有道上线的 YNMT 技术，由网易有道与网易杭州研究院历时两年合力研发，让以中文为中心的、根据中文用户使用习惯定制的神经翻译系统服务于 6 亿有道用户，服务于有道词典、有道翻译官、有道翻译网页版、有道 e 读等产品。

除了 BAT 这类大型的 IT 公司，一些机器翻译的创业公司如火如荼的发展起来。例如“小牛”翻译，由东北大学计算机科学与工程学院自主研发的机器翻译系统 Niu Trans，荣获钱伟长中文信息处理科学技术一等奖，这是国内中文信息处理领域的最高科学技术奖项。小牛翻译是目前国际上功能最强的两个开源统计机器翻译之一，目前有 70 多个国家的 2000 多个高校和企业研究机构下载使用。它不仅能翻译外文，还能翻译西藏、新疆等少数民族语言。由微软亚洲研究院和微软搜索技术中心的资深技术专家创立的爱特曼科技（Atman）是一家

人工智能创业公司，创立仅三个月产品还没上线便获得千万级的天使轮投资。该公司聚焦于世界领先机器翻译技术的研发和应用，核心技术有机器翻译、语音识别、机器写作、知识图谱等，提供的产品和服务包括：领先机器翻译技术结合译后编辑重构高质量语言转换服务、外媒内容全链条生产平台，包括外媒选材、机器翻译、在线编辑、自动分发等。

总而言之，机器翻译在我国从无到有，现如今其发展更是有着新的广度和深度，深刻的时代意义和现实价值。

2 technology

技术篇



2 技术篇

机器翻译的过程包括三个阶段，原文分析、原文译文转换和译文生成。

根据不同的翻译目的和翻译需求，在某一具体的机器翻译系统中，可以将原文分析和原文译文转换相结合，独立出译文生成，建立相关分析独立的生成系统。在这一翻译过程当中，机器翻译在进行原文分析时要考虑文本的结构特点，而在译语生成时则不考虑源语的结构特点。也可以结合原文译文转换与译文生成，把原文分析独立出来，建立独立分析相关生成系统。此时，文本分析时不考虑译语的结构特点，而在译语生成时要考虑源语的结构特点。还可以让原文分析、原文译文转换与译文生成分别独立，建立独立分析独立生成系统。在这样的系统中，分析源语时不考虑译语的特点，生成译语时也不考虑源语的特点，通过原文译文转换解决源语译语之间的异同。

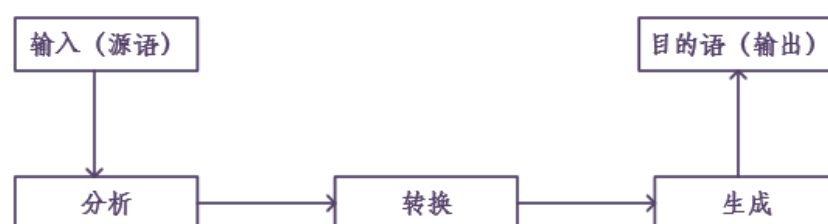


图 3 机器翻译过程

自机器翻译诞生以来，其研究围绕理性主义方法和经验主义方法两种思潮进行了两次转变。

所谓“理性主义”的翻译方法，是指由人类专家通过编撰规则的方式，将不同自然语言之间的转换规律生成算法，计算机通过这种规则进行翻译。这种方法理论上能够把握语言间深层次的转换规律，然而理性主义方法对专家的要求极高，不仅要求了解源语言和目标语言，还要具备一定的语言学知识和翻译知识，更要熟练掌握计算机的相关操作技能。这些因素都使得研制系统的成本高、周期长，面向小语种更是人才匮乏非常困难。因此，翻译知识和语言学知识的获取成为基于理性的机器翻译方法所面临的主要问题。

所谓“经验主义”的翻译方法，指的是以数据驱动为基础，主张计算机自动从大规模数据中学习自然语言之间的转换规律。由于互联网文本数据不断增长，计算机运算能力也不断加强，以数据驱动为基础的统计翻译方法逐渐成为机器翻译的主流技术。但是同时统计机器翻译也面临诸如数据稀疏、难以设计特征等问题，而深度学习能够较好的环节统计机器翻译所面临的挑战，基于深度学习的机器翻译现在正获得迅速发展，成为当前机器翻译领域的热点。

2.1 理性主义方法

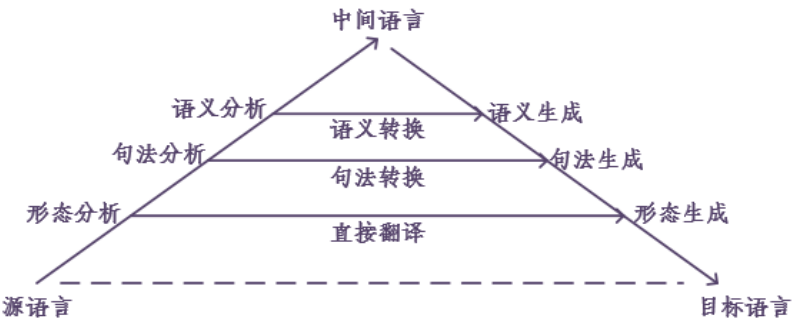
2.1.1 基于规则的机器翻译

基于规则的机器翻译方法（Rule-based System）的基本思想认为，一种语言无限的句子可以由有限的规则推导出来。依据语言规则对文本进行分析，再借助计算机程序进行翻译，这是多数商用机器翻译系统采用的方法。

基于规则的方法比较直观，能够直接表达语言学家的知识。规则的颗粒具有很大的可收缩性，大颗粒度的规则具有很强的概括能力，而且有比较好的系统适应性，不依赖于具体的训练语料；小颗粒度的规则具有精细的描述能力，这种方法便于处理复杂的结构和进行深层次的理解，如解决长距离依赖等问题。

但是，基于规则的翻译方法中规则主观因素比较重，有时与客观事实有一定差距；规则的覆盖性比较差，特别是细颗粒度的规则很难总结得比较全面；规则之间的冲突没有好的解决办法；规则库的调试是一个漫长枯燥的过程；规则一般只局限于某一个具体的系统，规则库开发成本太高。

图 4 机器翻译的转换层面



基于规则的机器翻译系统中，主要包括词法、句法、短语规则和转换生成语法规则，通过三个连续的阶段实现分析、转换、生成，根据三个阶段的复杂性可以分为直接翻译、结构转换翻译和中间语翻译。

(1) 直接翻译

直接翻译是指把源语中的单词或句子直接替换成相应的目的语的单词，必要时可以对词序进行适当的调整。这是机器翻译最初构想的体现，从目的语中寻找与源语词汇相对应的单词，但并不是电子词典 word-to-word 的形式，而是翻译句子中的所有词汇，再通过词语翻译、插入、删除和局部的词序调整来实现翻译，不进行深层次的句法和语义的分析，直接翻译应用的后期也加入了一些简单的句法或者是语义规则，对替换后的词语进行重新排序，生成最终的目的语文本，也可以采用一些统计方法对词语和词类序列进行分析。

直接翻译是早期机器翻译系统常用的方法，后来 IBM 提出的统计机器翻译模型也可以认为是采用了这一范式，著名的机器翻译系统 Systran 早期也是采用这种方法，后来逐步引入了一些句法和语义分析。

由于目的语和源语在句子语法结构等方面的差别很大，所以使用直接翻译法翻译出来的句子可读性和准确性都比较低，但它是机器翻译最实质性的一步，是机器翻译变成现实的一次迈步。



图 5 直接翻译过程

(2) 结构转换翻译

结构转换翻译是在直接翻译系统上出现的,相比较于直接翻译,它更多的从句子的层面来分析处理源语与目的语,译文的可读性和准确性更高。结构转换翻译通常包括分析、转换和生成三个阶段。分析要对源语言句子和源语言深层结构进行分析,其中相关分析在分析时要考虑目标语言的特点,而独立分析在分析过程中则与目标语言无关。从源语深层结构向目标语言的深层结构转换是关键部分,生成则是由目标语言深层结构生成目标语言句子,相关生成要考虑语言的特点,独立生成则与源语言无关。这种方法被认为是模拟人类翻译活动最恰当的机制。不同的语言具有相同或者相似的深层结构,就像是一座桥梁,把人类不同的语言连接起来,使得两种语言间可以实现翻译交流。目前绝大部分商品化机器翻译系统采用转换式机器翻译方法。

理想的转换方法应该做到独立分析和独立生成,这样在进行多语言翻译的时候可以大大减少分析和生成的工作量;转换放大根据深层结构所处的层面可分为句法层转换和语义层转换,分别对应句法信息和语义信息;分析的深层次越深,歧义排除也就越充分,但同时,错误率也会相对越高。

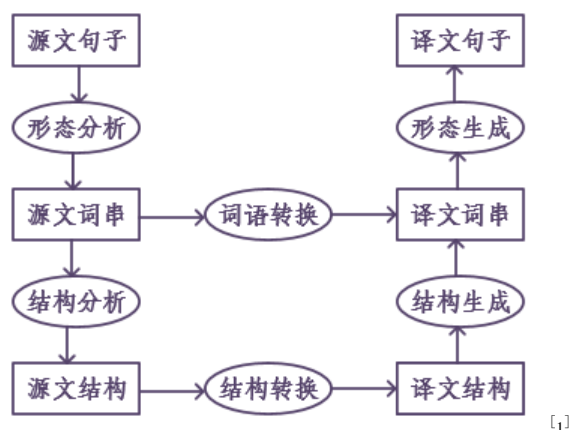


图 6 基于转换方法的翻译流程

人类自然语言中很多单词不止有一个意思,比如中文的“意思”二字就有很多不同的意思,容易产生歧义。在机器翻译中,为了简化比较复杂的表达结构,避免翻译过程中出现有歧义的语言现象,能够独立于各种自然语言,同时还能够清晰准确的表达各种自然语言的人造计算机语言便应运而生,这种作为翻译中介的人造计算机语言被称作中间语。它常见的形式有语义网络 (Semantic Network)、框架 (Frame) 和逻辑 (Logic), 以某种知识表示形式作为中间语言的机器翻译方法有时候也称为基于知识的机器翻译方法。

(3) 中间语言转换翻译

中间语言转换的机器翻译原理其实是在不同的语言之间建立一个通用的语义-句法表达式。整个翻译过程分为“分析”和“生成”两个阶段,由源语言到中间语言的生成,由中间语言到目标语言的生成环节。分析过程只与源语言有关,与目标语言无关,生成过程只与目标语言有关,与源语言无关。

[1] 刘群 机器翻译原理与方法讲义

中间语言方法的优点在于进行多语种翻译的时候，只需要对每种语言分别开发一个分析模块和一个生成模块，模块总数为 $2*n$ ，相比之下，如果采用转换方法就需要对每两种语言之间都开发一个转换模块，模块总数为 $n*(n-1)$ 。

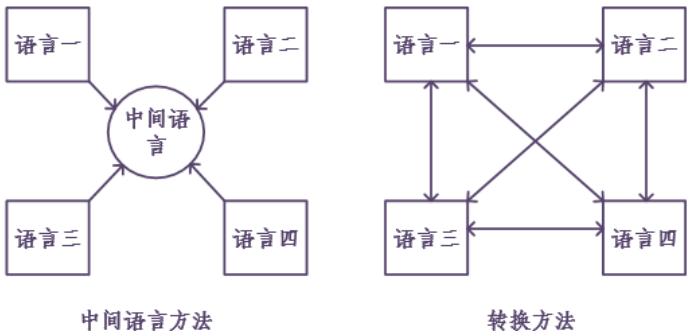


图 7 中间语言与转换方法比较

中间语言方法一般用于多语言的机器翻译系统中，从实践看，采用某种人工定义的知识表示形式作为中间语言进行多语言机器翻译都不太成功，如日本主持的亚洲五国语言机器翻译系统，总体上是失败的。在 CSTAR 多国机器翻译系统中，曾经采用了一种中间语言方法，其中间语言是一种带话语信息的语义表示形式，由于语音翻译都限制在非常狭窄的领域中（如旅游领域或机票预定），语义描述可以做到比较精确，因此采用中间语言方法有一定的合理性，但该方法最终也不成功。

实际上，领域特别窄的场合可以采用中间语言方法，一个适合于中间语言方法的例子是数词的翻译，采用阿拉伯数字作为中间语言显然是比较合理的。

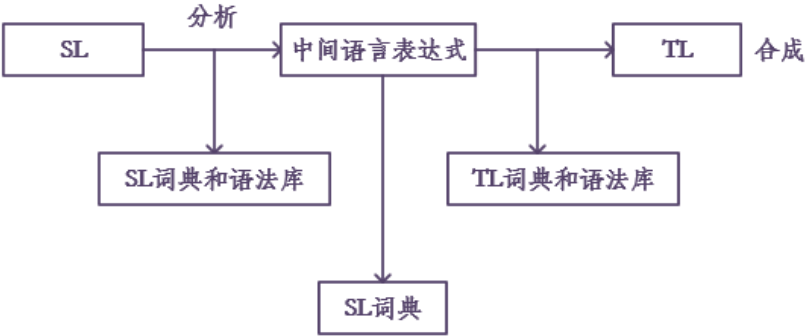


图 8 中间语转换翻译过程

2.2 经验主义方法

20 世纪 80 年代末至 90 年代初，随着计算机技术的快速发展，大规模双语语料库的构建以及机器学习方法的兴起，机器翻译方法逐渐由基于理性主义思维的规则方法转向基于经验主义思维的语料库方法。基于语料库的机器翻译方法又可以进一步划分为基于实例的翻译方法和基于统计模型的翻译方法。基于语料库的方法使用语料库作为翻译知识的来源，无需人工编写规则，系统开发成本低，速度快；而且从语料库中学习到的知识比较客观，覆盖性也比较好的。但是这种系统性能严重依赖于语料库，有着严重的数据稀疏问题，也不容易获得大颗粒度的高概括性知识。

2.2.1 基于统计的机器翻译

统计机器翻译（Statistics-based machine translation）的基本思想是充分利用机器学习技术，通过对大量的平行语料进行统计分析进行翻译。通俗来讲，源语到目的语的翻译过程是一个概率统计的问题，任何一个目的语句子都有可能是任何一个源语的译文，只是概率不同，机器翻译的任务就是找到概率最大的那个句子。

20 世纪 90 年代初期，IBM 的研究人员提出了基于信源信道思想的统计机器翻译模型，并在实验中获得了初步的成功，正式标志着统计机器翻译时代的到来。不过由于当时计算机能力等方面限制，真正展开机器翻译方法研究的人并不多，统计机器翻译方法是否有效还受到人们的普遍怀疑，随着越来越多的人员投入到统计机器翻译中并取得成功，统计方法已经逐渐成为国际上机器翻译研究的主流方法之一。

最初 IBM 研究人员提出的是基于词的机器翻译模型，但是，由于这种机器翻译模型复杂度较高，翻译质量也不尽人意，因此逐渐被一些更加有效的翻译模型所替代。下图是当前机器翻译中一些典型的翻译模型。

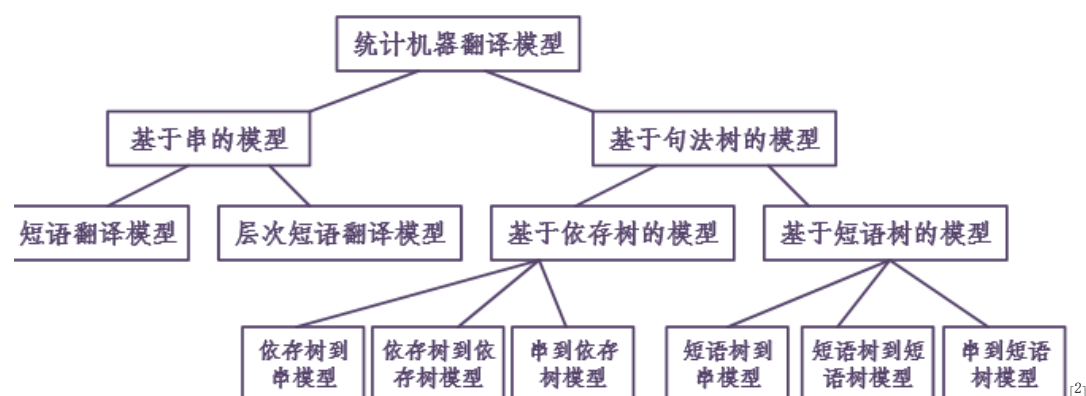


图 9 统计机器翻译典型模型

统计机器翻译也是基于语料库的机器翻译方法，不需要人工撰写规则，而是从语料库中获取翻译知识，这一点与基于实例的方法相同。为翻译建立统计模型，把翻译问题理解为搜索问题，即从所有可能的译文中选择概率最大的译文，基于实例的机器翻译则无需建立统计模型，二者的区别还在于，基于实例的机器翻译中，语言知识表现为实例本身，而统计机器翻译中，翻译知识表现为模型参数。

统计机器翻译是以严格的数学理论做基础的。所有的翻译知识都是以概率的形式呈现，表现为某种参数。训练的过程就是为了得到这些参数，解码的过程则是利用这些参数去搜索匹配最好的译文，只要使用这些参数就不需要去搜索原始的语料库。在整个过程中，机器翻译并不需要人工构造的翻译知识，所有的语言知识都是从语料库中自动获取。统计机器翻译的成功在于采用了一种新的研究范式，这种研究范式已在语音识别等领域中被证明是一种成功的翻译，但在机器翻译中是首次使用。这种范式的明显特点是，公开的大规模的训练数据、周期性的公开测评和研讨以及开放源码工具。

^[2]机器翻译原理与方法讲义

目前，统计机器翻译所使用的语料库是双语句子对齐的语料库，规模通常在几万句对到几百万句对不等。统计机器翻译的过程被看作是一个最优解搜索的过程，系统从巨大的可能译文中搜索最优的译文，搜索所使用的算法则采用人工智能中的一些成熟算法。

统计翻译模型的发展，迄今为止经历了三个阶段。分别是基于词的模型，基于短语的模型和基于句子的模型。基于短语的模型中的“短语”表示连续的词串，该模型的基本思想是：首先从双语句子对齐的平行语料库中抽取短语到短语的翻译规则，在翻译时将源语言句子切分为短语序列，利用翻译规则得到目标语言的短语序列，然后借助调序模型对目标语言短语序列进行排序，最终获得最佳的目标译文。其中，短语调序模型，尤其是长距离的短语调序，一直是短语翻译模型的关键问题。目前，基于短语的模型是最为成熟的模型，而基于句子的模型是当前研究的热点。统计机器翻译的模型可以表现为一个金字塔的形式，如图 10。

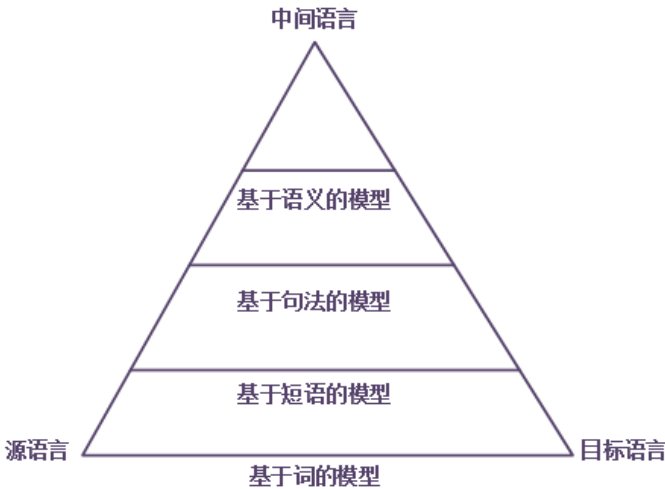


图 10 基于统计的机器翻译模型

在这个金字塔上，越往塔尖的方向走，对语言的分析也越深入。理论上来说，对语言的分析越深入，所具有的排歧能力就应该越强，译文的质量也应该越高。但实际上，分析语言本身就是一个很难的问题，分析的深度越深，往往引入的错误也越多，反而会导致翻译质量的下降。因此，如何通过引入更深层的语言分析来提高模型的排歧能力，同时又要避免分析导致的错误，就成了统计翻译模型要解决的主要问题。

统计机器翻译为自然语言翻译过程建立概率模型并利用平行语料库训练模型参数，无需人工编写规则，利用语料库直接训练得到机器翻译系统，人工成本低、开发周期短，只要有语料库就很容易适应新的领域或者语种，成为 Google、微软、百度等国内外公司在线翻译系统的核心技术。

尽管如此，统计机器翻译仍然面临着一些严峻的挑战。例如统计机器翻译依赖人类专家通过特征来表示各种翻译知识源，由于语言之间的结构转换非常复杂，人工设计特征难以保证覆盖所有的语言现象；统计机器翻译中的原规则结构复杂，对语料库的依赖性强，引入复杂的语言知识比较困难，即使现在可以用大规模语料库训练数据，但仍然面临着严重的数据稀疏问题。

2.2.2 基于实例的机器翻译

基于实例的翻译方法（Example-based Machine Translation）由日本翻译专家长尾真（Makoko Nagao）提出，他在 1984 年发表了《采用类比原则进行日-英机器翻译的一个框架》一文，探讨日本人初学英语时翻译句子的基本过程，长尾真认为，日本人初学英语时总是记住一些最基本的英语句子以及一些相对应的日语句子，他们要对比不同的英语句子和相对应的日语句子，并由此推论出句子的结构。参照这个学习过程，在机器翻译中，如果我们给出一些英语句子的实例以及相对应的日语句子，机器翻译系统可以通过识别和比较这些实例以及译文的相似之处和相差之处，从而挑选出正确的译文。

在基于实例的机器翻译系统中，系统的主要知识源是双语对照的翻译实例库，实例库主要有两个字段，一个字段保存源语言句子，另一个字段保存与之对应的译文，每输入一个源语言的句子时，系统把这个句子同实例库中的源语言句子字段进行比较，找出与这个句子最为相似的句子，并模拟与之相对应的译文，最后输出译文，这是一种由实例引导推理的机器翻译方法，整个翻译过程其实是查找和复现类似的例子，不需要对源语言进行任何分析，只需要通过类比，发现和记起特定的源语言表达或以前的翻译实例作为主要知识源来对新的句子进行翻译。

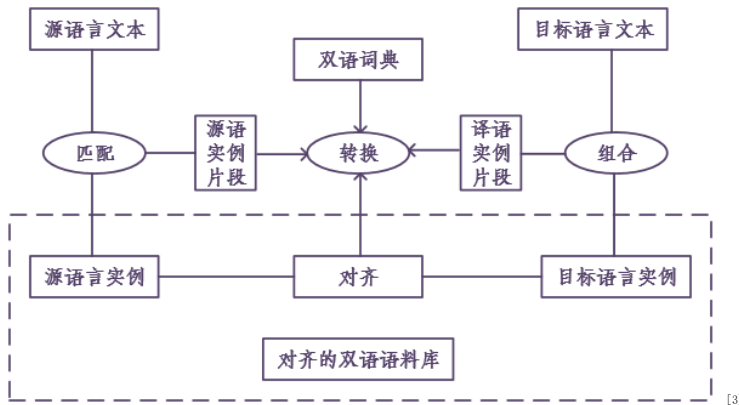


图 11 基于实例方法翻译过程

基于实例的机器翻译系统中，翻译知识以实例和语义类词典的形式表示，易于增加或删除，系统的维护简单易行，且利用了较大的翻译实例库并进行精确地对比，有可能产生高质量译文，而且避免了基于规则的那些传统的机器翻译方法必须进行深层语言学分析的难点，在翻译策略上很有吸引力的。

基于实例的机器翻译直接使用对齐的语料库作为知识表示形式，知识库的扩充非常简单，而且不需要进行深层次的语言分析，也可以产生高质量的译文。

但是基于实例的机器翻译系统的翻译质量取决于翻译记忆库的规模和覆盖率，至少要百万句对以上，因此如何构建大规模翻译记忆库成为影响基于实例的机器翻译研究的关键。现阶段，由于缺少大规模的双语对齐语料库，基于实例的机器翻译系统匹配率其实并不高，往往只有限定在特定的专业领域时，翻译效果才能达到使用要求。如果基于实例机器翻译匹配

[3]机器翻译原理与方法讲义

成功，可以获得相对较高质量的译文，因此基于实例的机器翻译一般和基于规则的机器翻译相结合使用，会产生比较好的翻译结构。对于匹配率过低的问题，可以试着做到短语级别的双语对齐，以提高匹配命中率，通过短语级别的局部匹配，结合相应的目标句子的框架，完成句子的翻译。

2.2.3 基于深度学习的机器翻译

从最初的基于规则的机器翻译到最新的依靠数据驱动进行的机器翻译，其总体发展趋势是要让计算机更加自主的学习如何翻译。利用平行语料库进行数据的训练，是提高机器翻译准确性和可读性的关键，深度学习的引入则成了当前热点。

(1) 深度学习发展脉络

以下是 AMiner 研究人员通过四个脉络对深度学习发展脉络进行了梳理。

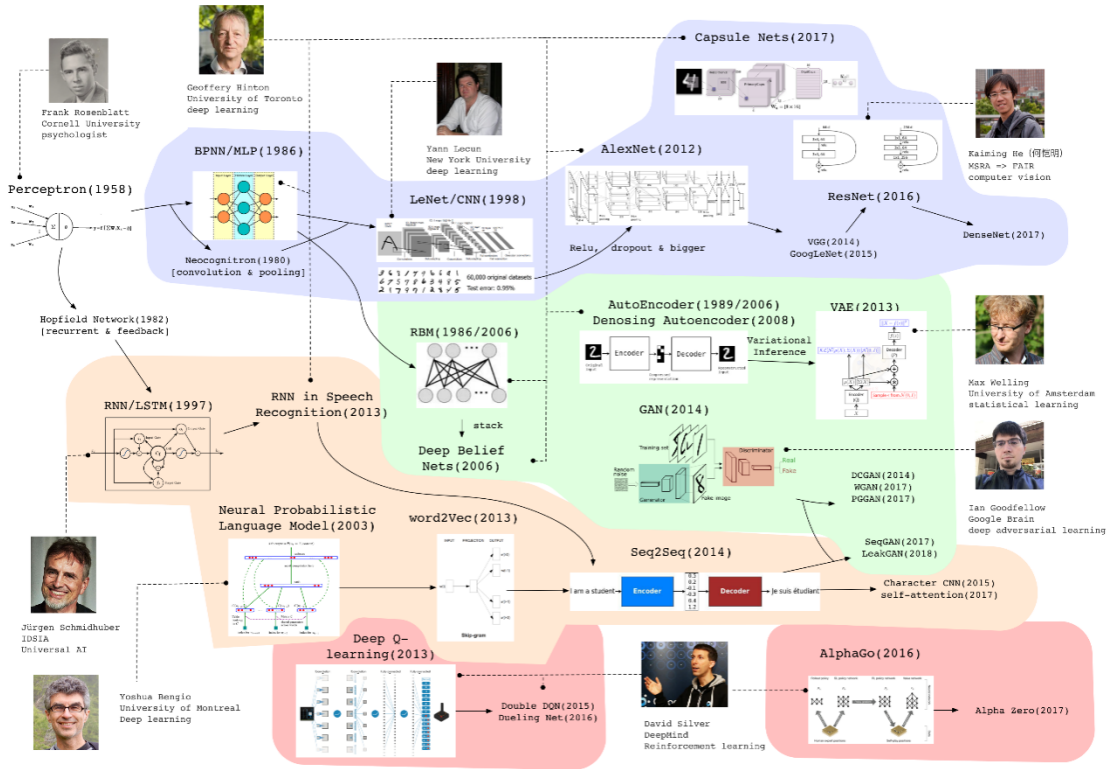


图 12 深度学习发展脉络

● 脉络一 cv/tensor

1943 年卡洛可和皮茨提出了抽象的神经元模型 MP，该模型可以看作深度学习的雏形。1957 年 Frank Rosenblatt 发明了感知机，是当时首个可以学习的人工神经网络。1969 年 Marvin Minsky 和 Seymour Papert 用详细的数学证明了感知机的弱点，神经网络研究进入冰河期。1984 年福岛邦彦提出了卷积神经网络的原始模型神经感知机，产生了卷积和池化的思想。1986 年 Hinton 等人提出一般 Delta 法则，并用反向传播训练 MLP。1998 年以 Yann LeCun 为首的研究人员实现了 5 层的卷积神经网络——LeNet-5，以识别手写数字。LeNet-5 标志着 CNN（卷积神经网络）的真正面世，LeNet-5 的提出把 CNN 推上了一个小高潮。

之后 SVM 兴起。2012 年 AlexNet 在 ImageNet 上夺冠，掀起了深度学习的热潮。AlexNet 可以算是 LeNet 的一种更深更宽的版本，并加上了 relu、dropout 等技巧。这条思路被后人发展，出现了 VGG，GoogLeNet 等网络。2016 年何恺明在层次之间加入跳跃连接，Resnet 极大增加了网络深度，效果有很大提升。cvpr best paper densenet 也是沿着这条思路发展的。

除此之外，cv 领域的特定任务还出现了各种各样的模型（Mask-RCNN 等），这里不一一介绍。2017 年 Hinton 认为反省传播和传统神经网络有缺陷，继而提出了 Capsule Net。但是目前在 cifar 等数据集上效果一般，这条思路还需要继续验证和发展。

● 脉络二 生成模型

传统的生成模型是要预测联合概率分布 $P(x, y)$ 。RBM 本在 1986 年的时候就存在，只是 2006 年重新作为一个生成模型，并且堆叠成为 deep belief network，使用逐层贪婪或者 wake-sleep 的方法训练，Hinton 等人从此开始使用深度学习重新包装神经网络。

Auto-Encoder 提出于上世纪 80 年代，现在随着计算能力的进步重新登上舞台。2008 年，Bengio 等人又提出 denoise Auto-Encoder。Max Welling 等人使用神经网络训练 Variational auto-encoder。此模型可以通过隐变量的分布采样，经过后面的 decoder 网络直接生成样本。

GAN（生成对抗网络）于 2014 年提出。它是一个生成模型，通过判别器 D 和生成器 G 的对抗训练，直接使用神经网络 G 隐式建模样本整体的概率分布。每次运行便相当于从分布中采样。DCGAN 是较好的卷积神经网络实现，而 WGAN 则是通过维尔斯特拉斯距离替换原来的 JS 散度来度量分布之间的相似性工作，训练更稳定。PGGAN 则逐层增大网络，生成极其逼真的人脸。

● 脉络三 Sequence Learning

1982 年出现的 Hopfield Network 有了递归网络的思想。1997 年 Schmidhuber 发明 LSTM，并做了一系列的工作。但是更有影响力的还是 2013 年由 Hinton 组使用 RNN 做的语音识别工作。

文本方面，Bengio 提出了一种基于神经网络的语言模型，后来 Google 提出 word2vec 也包含了一些反向传播的思想。在机器翻译等任务上，逐渐出现了以 RNN 为基础的 seq2seq 模型，模型通过编码器把一句话的语义信息压成向量再通过解码器输出，但更多的还要和注意力模型结合。之后以字符为单位的 CNN 模型在很多语言任务也表现不俗，而且时空消耗更少。LSTM/RNN 模型中的注意力机制是用于克服传统编码器-解码器结构存在的问题的。其中，自注意力机制实际上就是采取一种结构令其同时考虑同一序列局部和全局的信息。

● 脉络四 Deepreinforcement Learning

该领域最出名的是 DeepMind，这里列出的 David Silver 则是一直研究 reinforcement learning（rl，强化学习）的高管。

q-learning 是很有名的传统 rl 算法，deep q-learning 则是将原来的 q 值表用神经网络代替。之后 David Silver 等人又利用其测试了许多游戏，发在了 Nature 上。增强学习在 double duel 的进展，主要是 q-learning 的权重更新时序。DeepMind 的其他工作诸如 DDPG、A3C 也非常有名，它们是基于 policy gradient 和神经网络结合的变种。

可以说基于深度预训练的机器翻译，显著地提升了机器翻译的质量，接近普通人的水平，是当前机器翻译领域的热点。大致可以分为两种情况，一是沿用深度学习改进统计机器翻译中的相关模块；二是直接利用神经网络实现源语言到目标语言的映射，即端到端的神经机器翻译。

（2）利用深度学习改进统计机器翻译

利用深度学习改进统计机器翻译是指利用深度学习改进统计机器翻译中的相关模块，如语言模型、翻译模型等。上文也提到统计机器翻译有着不可避免的缺点，其中较为严重的是数据稀疏问题，而深度学习可以帮助统计翻译模型较好的解决这一问题。机器翻译的核心是语言模型，语言模型对译文的流利度和质量都有着至关重要的作用，通过深度学习可以改进语言模型。**n-gram** 是传统的语言模型所采用的方法，模型参数是通过极大似然估计训练所得，采用离散表示（每个词都是独立的符号），但是因为大多数 **n-gram** 在语料库中只出现一次，无法准确估计模型参数，所以极大似然估计面临着严重的数据稀疏问题。因此传统的统计机器翻译基本会使用平滑和回退等策略来缓解数据稀疏问题，但即使采用平滑和回退策略，统计机器翻译还是因为数据稀疏无法获得更多历史信息，通常仅能使用 **4-gram** 或者 **5-gram** 语言模型。

深度学习著名代表学者 **Yoshua Bengio** 教授 2003 年提出基于神经网络的语言模型，这一模型中的数据稀疏问题由于分布式表示的存在得到了有效缓解；2014 年美国 **BBN** 公司的研究人员进一步提出神经网络联合模型（**Neural Network Joint Models**）。他们提出，对于决定当前词来说，不仅仅是目标语言端的历史信息有着重要的作用，源语言端的相关部分也起着很重要的作用。这种观点是对传统的语言模型的一种颠覆，因为传统的语言模型往往只考虑目标语言端的前 **n-1** 个词，而不会探索更多。使用分布式表示能够缓解数据稀疏问题，再加上神经网络联合模型能够使用丰富的上下文信息，融入了深度学习的翻译方法相对应传统的翻译方法有了一个质的进步。

使用神经网络对机器翻译来说，还有一个更为显著的好处，即能够解决特征难以涉及的问题。例如调序模型。基于短语的统计机器翻译的重要调序方法之一是基于反向转录文法的调序模型。这种模型将调序视作一个二元分类问题，即两个相邻源语言词串的译文顺序有顺序拼接和逆序拼接两种处理方法。传统方法通常使用最大熵分类器，但是这一方法面临着一个重要难点，即如何设计能够捕获调序规律的特征。要从众多的词语集合中选出能够对调序决策起到关键作用的词语是非常困难的，而且词串的长度一般都非常长，更是为此增加了难度。所以，基于反向转录文法的调序模型由于无法把握基于词串的特征设计问题，从而无法充分利用整个词串的信息，造成了信息的白白流失。利用神经网络能够较好的缓解特征设计的问题，首先词串的分布式表示可以利用递归自动编码器生成，然后神经网络分类器可以基于四个词串的分布式表示来建立。因此，基于神经网络的调序模型不需要人工参与设计特征就能够自主利用整个词串的信息，调序分类准确率和翻译质量显著提高。实际上，深度学习不仅停留在为机器翻译生成新的特征这一步，更能够将现有的特征集合转化生成新的特征集合，翻译模型的表达能力有了显著提升。

（3）端到端神经机器翻译

端到端神经机器翻译（**End-to-End Neural Machine Translation**）是一种全新的机器翻译

方法,于 2013 年兴起。这种翻译方法通过神经网络直接将源语言文本映射成目标语言文本。这种方法仅通过非线性的神经网络便能直接实现自然语言文本的转换,不再需要由人工设计词语对齐、短语切分、句法树等隐结构,也不需要人工设计特征。

2013 年,英国牛津大学的 Nal Kalchbrenner 和 Phil Blunsom 提出了端到端的神经机器翻译,他们提出了一个新框架,即“编码-解码”的框架。对于一个源语言句子,首先将它映射为一个连续、稠密的向量,这一过程是通过编码器实现的,然后再将这个向量转化为目标语言的句子,这一过程通过解码器实现。Kalchbrenner 和 Blunsom 在论文中使用的编码器是卷积神经网络(Convolutional Neural Network),解码器是递归神经网络(Recurrent Neural Network)。能够捕获全部历史信息和处理变长字符串是递归神经网络的优点。这种新的架构,统计机器翻译的线性模型被非线性模型取代,隐结构流水线被单个复杂的神经网络取代,语义等价性通过连接编码器和解码器的向量来描述,同时,递归神经网络可以捕获无限长的历史信息。理论上,端到端的神经机器翻译能够捕获无限长的历史信息,可以取得理想的翻译效果,但是在真正处理长距离的依赖关系时还是有困难的。

为此,2014 年,长短期记忆(Long short-Term Memory)被美国 Google 公司引入端到端的神经机器翻译。通过设置门开关的方法长短期记忆能够较好的捕获长距离依赖。由于递归神经网络无论在编码器还是解码器里的使用,都使得端到端神经机器翻译性能得到了提升,取得了与传统机器翻译相当甚至更好地翻译效果。但是,要想实现准确的编码和语言句子,编码器都需要将它映射为一个维度固定的向量,这是一个极大的挑战。

Yoshua Bengio 研究组针对这一问题提出了基于注意力(Attention)的端到端神经网络翻译。其基本思想就是,在解码器生成单个的目标语言,有相关性的其实仅仅是小部分的源语言词,绝大多数源语言词都是无关的,这样就不需要使用整个源语言句子的向量,只要使用每个目标语言词相关的源语言端的上下文向量即可。为此,他们提出了一整套基于内容的注意力计算方法,这套方法能够更好地处理长距离依赖,同时提升端到端神经机器翻译的准确率和质量。

虽然端到端的神经机器翻译近年来发展迅速,但仍然存在很大的提升空间。首先是可解释性差。有别于传统的机器翻译在设计模型时根据语言学知识进行架构的方式,神经网络内部全部使用向量表示,从语言学的角度看可解释性很差,在设计新结构时如何融入语言学知识成为新的挑战。其次训练复杂度高,端到端神经机器翻译的训练复杂度是传统统计机器翻译不可比的,对于计算资源的依赖程度和要求也更高,想要获得较理想的实验周期必须使用较大规模的 GPU,因此计算资源成为端到端神经机器翻译的一个重要问题。

3 talent

人才篇



3 人才篇

机器翻译经历了几十年的发展，无数学者投身其中，国内外的研究学者甚多，本报告基于 AMiner 大数据，对该领域内的学者就行挖掘，并根据各学者在 AMiner 数据中的 H-index 排序选取其中几位进行简要介绍。

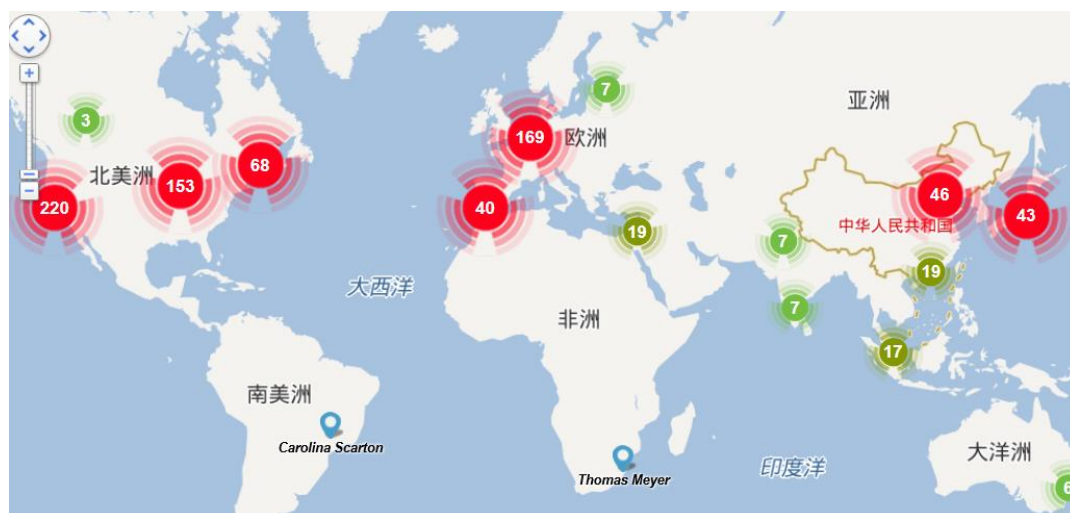


图 13 机器翻译领域全球学者分布

上图是以“machine translation”为关键词，在 AMiner 数据库得到的全球机器翻译领域人才分布图。由图可以看到，美国在这一领域人才最多且最为集中，欧洲和中国对机器翻译的研究紧跟其后，南美洲、非洲和大洋洲则人才相对匮乏。这与国家整体经济发展水平、教育水平、计算机水平有着密不可分的关系。



图 14 机器翻译领域中国学者分布图

我们以“machine translation”为关键字在 AMiner 数据库中对国内机器翻译领域人才进行挖掘，得到了国内机器翻译领域人才分布图。可以看出，机器翻译研究主要集中在北京，

这与北京高校众多、教育先进不无关系。

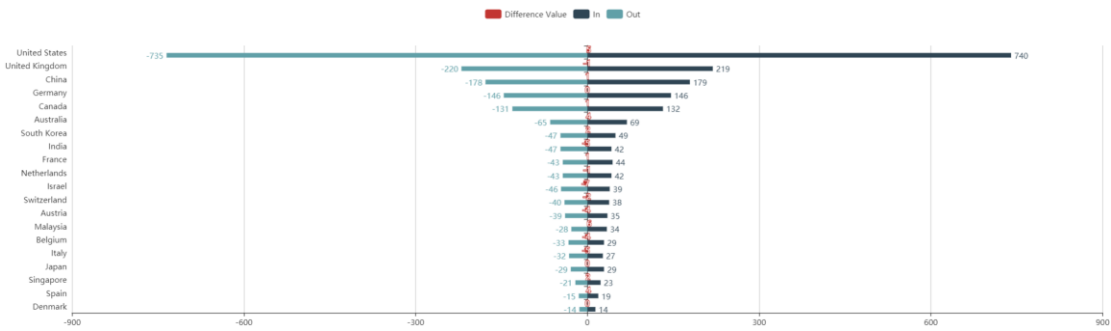


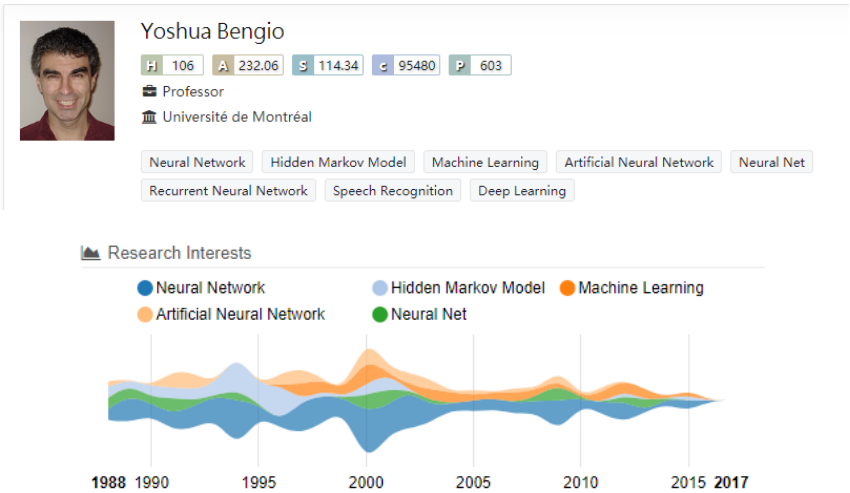
图 15 机器翻译各国人才顺逆差图

由上图可以看出，各国通信领域人才的流失和引进是相对比较均衡的，其中美国为通信人才流动大国，人才输入和输出幅度都大幅度领先。英国、中国、德国和加拿大等国落后于美国。其中，美国人才流动呈顺差趋势，英国、德国略微呈逆差趋势，中国和加拿大人才迁入迁出保持均衡。

AMiner 基于发表于国际期刊会议的学术论文，对某一领域内专家进行深度挖掘，并按照相关度和影响力等对专家进行排序和分类。排序规则参考 H-index、citation、论文数、专家所获得的荣誉、任职机构排名、专家活跃度、社交性及兴趣多样性等。多项指标综合分析，将该领域专家划分为领军人物、中坚力量和领域新星。

3.1 领军人物

● Yoshua Bengio



Yoshua Bengio 是加拿大蒙特利尔大学计算机科学与运筹学系的教授，机器学习实验室（MILA）的负责人。他的许多研究被广泛引用，其经典作品《Learning Deep Architectures for AI》非常适合对深度学习感兴趣的读者作为入门读物。

Bengio 的主要贡献在于他对 RNN 的一系列推动，包括经典的 neural language model, gradient vanishing 的细致讨论, word2vec 的雏形，以及现在的 machine translation；他也是神经网络复兴的主要的三个发起人之一。

Yoshua Bengio 是《Journal of Machine Learning Research》、《Neural Computation》和《Foundations and Trends in Machine Learning》的编辑。自 1999 年以来，一直与 Yann Le Cun 共同组织学习研讨会，并与他一起创建了国际代表性学习会议（ICLR）。

我们根据 AMiner 大数据，筛选出 Yoshua Bengio 发表论文中 citation 最高的几篇论文：

602	Deep learning Yann LeCun, Yoshua Bengio, Geoffrey Hinton Nature (2015) Cited by 7181 Bibtex https://doi.org/10.1038/nature14539	EI NATURE WOS
601	Learning Deep Architectures for AI Yoshua Bengio Foundations and Trends in Machine Learning (2009) Cited by 5131 Bibtex http://dx.doi.org/10.1561/22000000006	EI
600	A neural probabilistic language model Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Janvin Journal of Machine Learning Research (2000) Cited by 3259 Bibtex https://static.aminer.org/pdf/20170130/webConf/index.txt	EI
599	Representation learning: a review and new perspectives. Yoshua Bengio, Aaron Courville, Pascal Vincent IEEE Trans. Pattern Anal. Mach. Intell. (2013) Cited by 3036 Bibtex http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?tp=&arnumber=6472238	EI WOS

● Kevin Knight



Kevin Knight (武凯文)

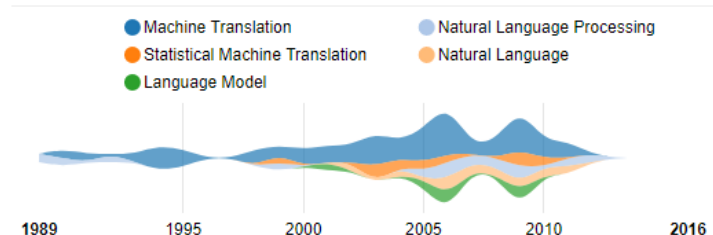
H 58 A 24.35 S 12.48 c 15881 P 154

Dean's Professor

Information Sciences Institute, University of Southern California

Machine Translation Natural Language Processing Statistical Machine Translation Natural Language
Language Model Translation Quality Word Alignment Knowledge Base

Research Interests



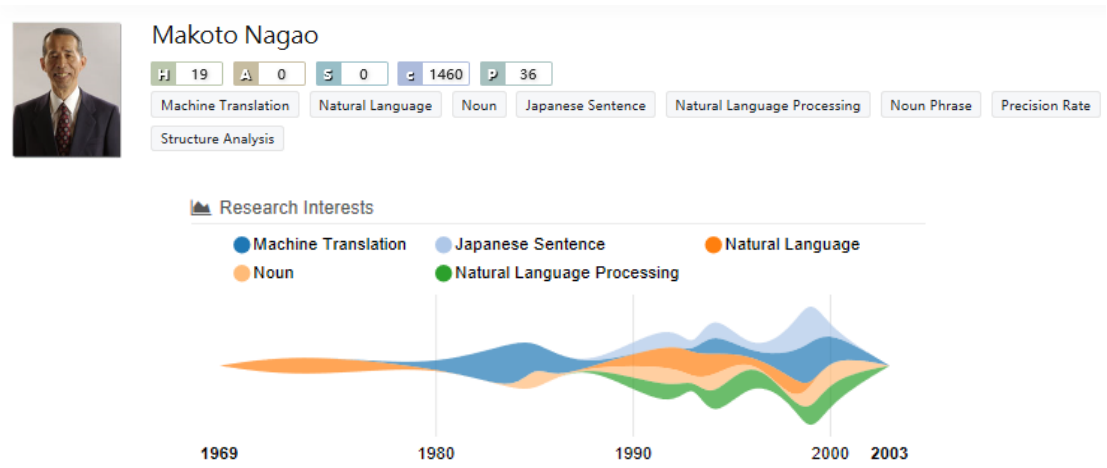
Kevin Knight 在卡内基梅隆大学计算机科学获得博士学位，目前是南加州大学信息科学研究所的一名教授，机器翻译界公认的领袖，他是统计机器翻译的主要倡导者之一，在统计机器翻译方面做了一系列的研究和推广工作，培养了一批知名学者，例如 Philipp Koehn 等，同时也是 JHU 的统计机器翻译夏季研讨班的主要组织者之一。

Kevin Knight 的研究兴趣包括自然语言生成器、自然语言处理、机器翻译、文本生成、密码学、人工智能、译码、计算语言学-自动语言翻译、自动文本摘要、大型词汇和分类法的构建以及与计算机的多媒体通信等。

我们根据 AMiner 大数据，筛选出 Kevin Knight 发表论文中 citation 最高的几篇论文：

154	Statistical Machine Translation phillipp koehn, kevin knight Encyclopedia of Machine Learning (2010) Cited by 1206 Bibtex http://www.statmt.org/book/	EI
153	Towards Distributed Use of Large-Scale Ontologies W. R. Swartout, R. Patil, K. Knight , T. Russ (1997) Cited by 912 Bibtex	
152	A syntax-based statistical translation model Kevin Knight , Kenji Yamada A syntax-based statistical translation model (2003) Cited by 851 Bibtex https://static.aminer.org/pdf/20160902/aclanthology/index.txt	EI
151	A syntax-based statistical translation model Kenji Yamada, Kevin Knight ACL (2001) Cited by 827 Bibtex http://www.aclweb.org/anthology/P01-1067	EI
150		

● NAGAO Makoto



长尾真是日本国家信息通讯研究院理事长，前京都大学校长。历任电子信息通讯学会会长，语言处理学会（ANLP）首任会长，信息处理学会会长，认知科学学会会长，日本机械翻译学会首任会长，国际图形识别联盟副会长，亚洲太平洋机械翻译协会（AAMT）首任会长，国际机械翻译学会（IAMT）首任会长。

他是国际自然语言处理与机器翻译领域杰出学科带头人，EBMT 之父，在图像和语言的智能信息处理等领域具有深厚的理论造诣并取得了丰硕的实践应用成果。

长尾真曾获得机械翻译国际联盟第一次荣誉奖章，国际人工智能财团学术研究奖，IEEE Emanuel R.Piore，信息处理学会论文奖，信息处理学会创立 20 周年纪念论文奖、功绩奖，电子信息通讯学会业绩奖，日本小林纪念特别奖、功绩奖，日本通商产业大臣表彰，日本邮政大臣表彰，日本科学技术信息中心丹羽奖，京都新闻文化奖，日本大川出版奖，日本紫绶奖章等。

我们根据 AMiner 大数据，筛选出长尾真发表论文中 citation 最高的几篇论文：

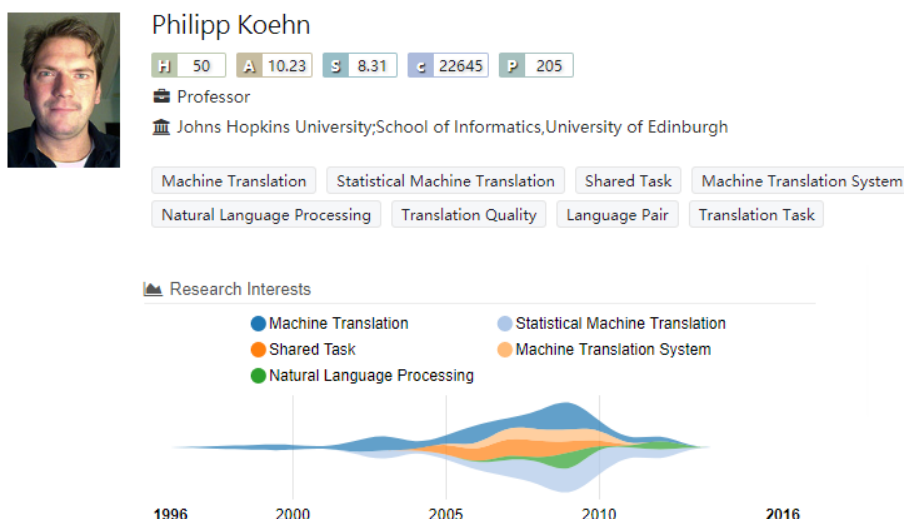
36 A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures
Sadao Kurohashi, Makoto Nagao
Computational Linguistics (1994)
Cited by 195 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

35 Building A Japanese Parsed Corpus
Sadao Kurohashi, Makoto Nagao
(2003)
Cited by 158 Bibtex http://dx.doi.org/10.1007/978-94-010-0201-1_14

34 A new method of N-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese
Makoto Nagao, Shinsuke Mori
COLING (1994)
Cited by 126 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

33 The Japanese government project for machine translation
Makoto Nagao, Jun-ichi Tsuji, Jun-ichi Nakamura
Computational Linguistics (1985)
Cited by 110 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

● Philipp Koehn



Philipp Koehn 毕业于南加州大学计算机科学系，目前是爱丁堡信息学院和约翰·霍普金斯大学计算机科学系的教授。

他的主要研究兴趣是统计机器翻译，并且是基于短语的机器翻译方法的发明者之一，这是 2013 年欧洲发明奖（EPO）研究的三个入围项目之一。基于短语的机器翻译方法是统计翻译方法的一个子领域，它使用一系列的词汇作为翻译的基础，扩展先前基于单词的方法。Philipp Koehn 指导并创建了摩西机器翻译译码器，这是一个开发统计机器翻译系统的平台，为很多语言对提供了一个平行的语料库，译码器是该领域研究的一个实际标准。

2003 年，他与 Franz Josef Och 和 Daniel Marcu 合著的论文《统计用语》在机器翻译界引起了广泛关注，并被引用三千多次。基于短语的方法被广泛应用于工业的机器翻译应用中，如谷歌翻译和亚洲在线。

我们根据 AMiner 大数据，筛选出 Philipp Koehn 发表论文中 citation 最高的几篇论文：

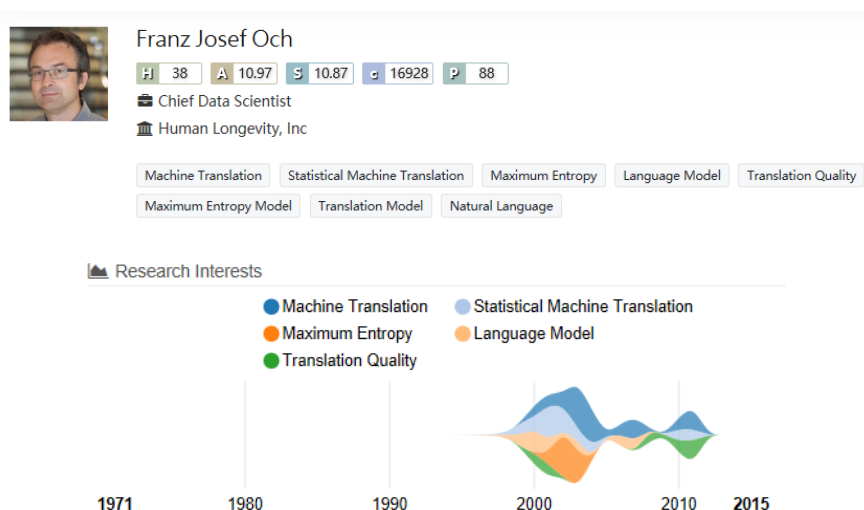
205
Moses: open source toolkit for statistical machine translation EI
Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens,
 ACL (2007)
 Cited by 4147 <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

204
Statistical phrase-based translation EI
Philipp Koehn, Franz Josef Och, Daniel Marcu
 HLT-NAACL (2003)
 Cited by 3309 <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

203
EuroParl: A Parallel Corpus for Statistical Machine Translation
Koehn Philipp
 (2005)
 Cited by 2307

202
Statistical Machine Translation EI
philipp koehn, kevin knight
 Encyclopedia of Machine Learning (2010)
 Cited by 1206 <http://www.statmt.org/book/>

● Franz Josef Och



Franz Josef Och 毕业于亚琛工业大学，是谷歌著名的研究科学家和谷歌翻译的首席架构师。目前统计机器翻译研究领域一些著名的开源软件，如 IBM 模型训练工具 Giza++、最大熵模型训练工具 YASMET 都是他开发的。在 Och 的研究中，数据规模总是第一位的。他也尝试过使用一些句法知识，但最后结论是，句法知识对统计机器翻译毫无用处，甚至有反作用。

他的主要研究包括统计机器翻译、自然语言处理和机器学习等方向。在机器翻译领域的主要贡献有：把判别模型引入机器翻译，从根本上取代 noisy-channel 模型而成为目前的标准模型框架；简化了基于短语的模型，Och 引入了相对频度，极大降低了参数估计的复杂度，这是 Och 的一个大贡献；开发并发布 GIZA++。他这些年来发表的很多论文，包括博士论文，都成了统计机器翻译研究领域的经典，被人广泛引用和验证。

我们根据 AMiner 大数据，筛选出 Franz Josef Och 发表论文中 citation 最高的几篇论文：

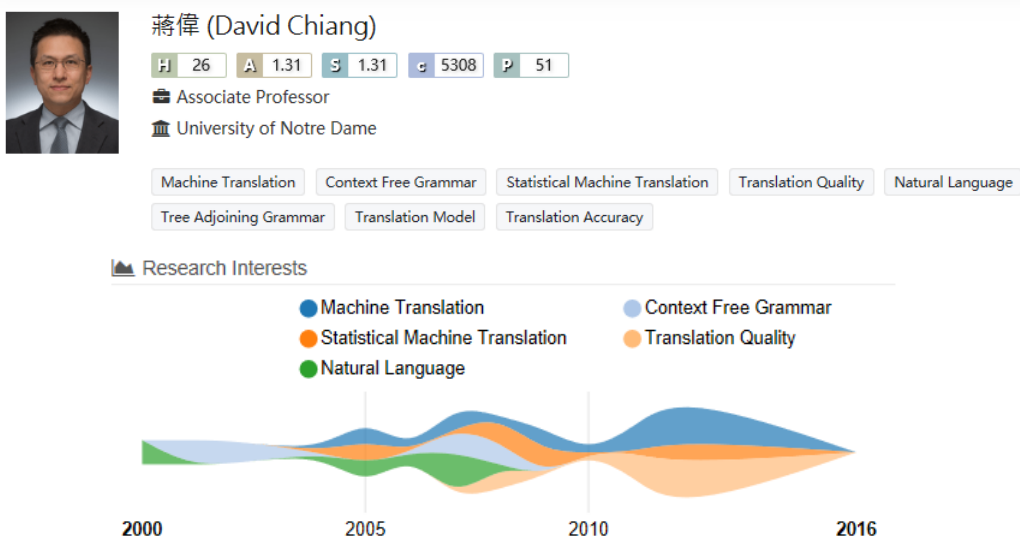
88 **A systematic comparison of various statistical alignment models** EI
 Franz Josef Och, Hermann Ney
 Computational Linguistics (2003)
 Cited by 3499 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

87 **Statistical phrase-based translation** EI
 Philipp Koehn, Franz Josef Och, Daniel Marcu
 HLT-NAACL (2003)
 Cited by 3309 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

86 **Minimum error rate training in statistical machine translation** EI
 Franz Josef Och
 ACL (2003)
 Cited by 1583 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

85 **Discriminative training and maximum entropy models for statistical machine translation** EI
 Franz Josef Och, Hermann Ney
 ACL (2002)
 Cited by 1153 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

● David Chiang



David Chiang 在宾夕法尼亚大学计算机与信息科学获得博士学位，目前是美国圣母大学的一名教授。

他的主要研究领域是自然语言处理，同时也很关注语言翻译，对句法分析等方面也很有研究。David Chiang 在 2005 年提出的基于短语的翻译模型，对机器翻译来说是一个巨大的进步，他把机器翻译从平面结构建模引向了层次结构建模。

我们根据 AMiner 大数据，筛选出 David Chiang 发表论文中 citation 最高的几篇论文：

51 **A hierarchical phrase-based model for statistical machine translation** EI
David Chiang
 ACL (2005)
 Cited by 1216 BibTeX <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

50 **Hierarchical Phrase-Based Translation** EI
David Chiang
 Computational Linguistics (2007)
 Cited by 1145 BibTeX <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

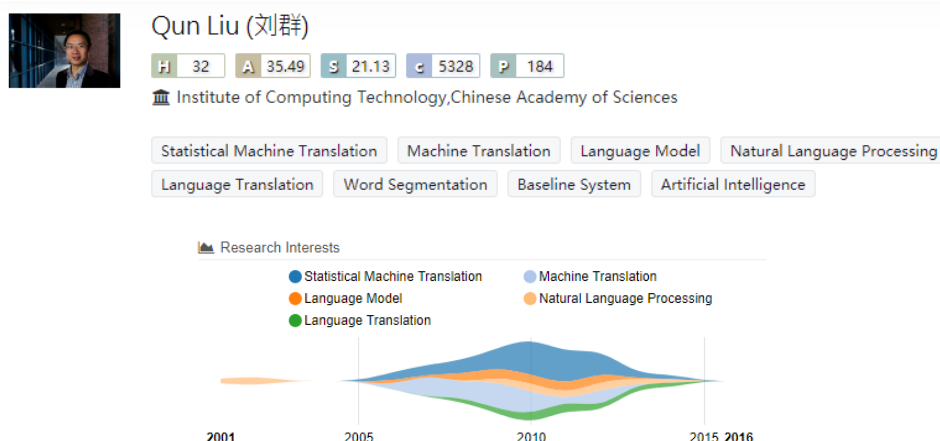
48 **Word Sense Disambiguation Improves Statistical Machine Translation** EI
 Yee Seng Chan, Hwee Tou Ng, **David Chiang**
 ACL (2007)
 Cited by 295 BibTeX <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

47 **Forest Rescoring: Faster Decoding with Integrated Language Models** EI
 Liang Huang, **David Chiang**
 ACL (2007)
 Cited by 277 BibTeX <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

3.2 中坚力量

以“machine translation”为关键字在 AMiner 数据库中对机器翻译领域一线中青年学者进行挖掘，找到相关度较高的学者包括厦门大学信息科学与技术学院史晓东教授、中科院自动化研究所宗成庆教授、东北大学自然语言处理实验室朱靖波教授、清华大学计算机科学与技术系刘洋教授、中国科学院计算技术研究所刘群教授、微软亚洲研究院周明老师、苏州大学计算机科学与技术系张敏教授、南京大学计算机科学与技术系陈家骏教授、哈工大计算机科学与技术学院赵铁军教授等。根据 H-index 排序，我们选取了前五位进行简要介绍。

● 刘群



刘群是中国科学院自然语言处理研究组组长，主要研究方向是中文自然语言处理，具体包括汉语词法分析、汉语句法分析、语义处理、统计语言模型、辞典和语料库、机器翻译、信息提取、中文信息处理和智能交互中的大规模资源建设和中文信息处理和智能交互中的评测技术等。

他曾负责 863 重点项目“机器翻译新方法的研究”和“面向跨语言搜索的机器翻译关键技术研究”等。

我们根据 AMiner 大数据，筛选出刘群发表论文中 citation 最高的几篇论文：

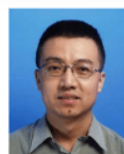
184
基於「知網」的辭彙語義相似度計算
劉群, 李袁建
(2002)
Cited by 1060 Bibtex

183
HHMM-based Chinese lexical analyzer ICTCLAS
Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, Qun Liu
SIGHAN (2003)
Cited by 497 Bibtex <https://aclanthology.info/papers/W03-1730/w03-1730>

182
 Tree-to-string alignment template for statistical machine translation
Yang Liu, Qun Liu, Shouxun Lin
ACL (2006)
Cited by 375 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

181
 Maximum entropy based phrase reordering model for statistical machine translation
Deyi Xiong, Qun Liu, Shouxun Lin
ACL (2006)
Cited by 290 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

● 张民



张民 (Min Zhang)

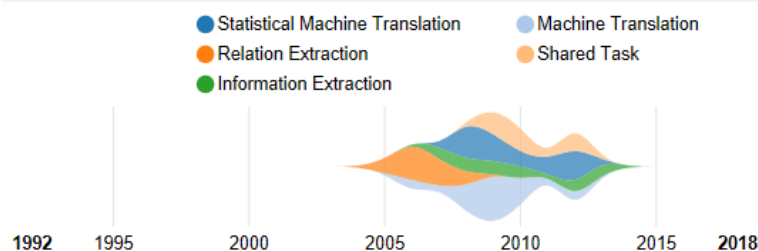
H 30 A 19.32 S 12.68 C 3531 P 129

Professor

苏州大学计算机科学与技术学院

Statistical Machine Translation Machine Translation Relation Extraction Information Extraction Shared Task
Machine Transliteration Common Benchmarking Platform Language Model



Research Interests







张民是苏州大学计算机科学与技术学院副院长。2003 年 12 月，他加入新加坡信息通信研究所并于 2007 年在研究所建立了统计机器翻译团队。2012 年加入苏州大学，并于 2013 年在该大学成立智能计算研究所。


他目前的研究兴趣包括机器翻译、自然语言处理、信息提取、社交网络计算、互联网智能、智能计算和机器学习。近年来在国际顶级学报和顶级会议发表学术论文 150 余篇，Springer 出版英文专著两部，主编 Springer 和 IEEE CPS 出版英文书籍十本。他一直积极地为研究界做贡献，组织多会议并在许多会议和讲座中进行演讲。

我们根据 AMiner 大数据，筛选出张民发表论文中 citation 最高的几篇论文：

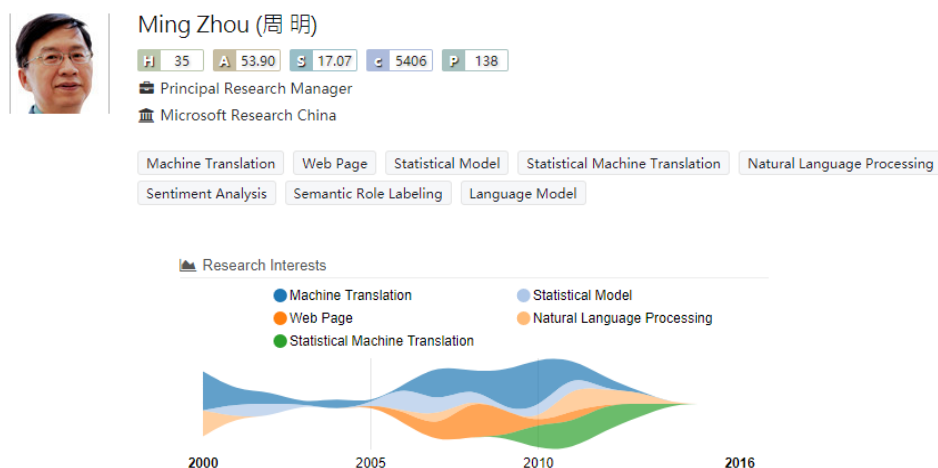
 A joint source-channel model for machine transliteration EI
 Haizhou Li, **Min Zhang**, Jian Su
 ACL (2004)
 Cited by 235  <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

 Tree Kernel-Based Relation Extraction with Context-Sensitive Structured Parse Tree Information EI SCOPUS WOS
 GuoDong Zhou, **Min Zhang**, Dong-Hong Ji, QiaoMing Zhu
 EMNLP-CoNLL (2007)
 Cited by 199  <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

 A Tree Sequence Alignment-based Tree-to-Tree Translation Model EI SCOPUS WOS
Min Zhang, Hongfei Jiang, AiTi Aw, Haizhou Li, Chew Lim Tan, Sheng Li
 ACL (2008)
 Cited by 117  <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

Report of NEWS 2009 machine transliteration shared task EI
 Haizhou Li, A. Kumaran, Vladimir Pervouchine, **Min Zhang**
 NEWS@IJCNLP (2009)
 Cited by 78  <https://aclanthology.info/papers/W09-3501/w09-3501>

● 周明



周明是微软亚洲研究院自然语言计算组的首席研究员和经理，是机器翻译和自然语言处理领域的专家。

他的研究兴趣包括搜索引擎、统计和神经机器翻译、问答、聊天机器人、计算机诗歌和文本挖掘等。

1989 年，他设计了“CEMT-I 机器翻译系统”，这是汉英机器翻译的第一个实验，获得了中国大陆政府的科学技术进步奖。1998 年，他设计了著名的中日文机器翻译软件产品 J-Beijing，并获得了日本机械翻译协会 2008 年颁发的机器翻译产品的最高荣誉称号。

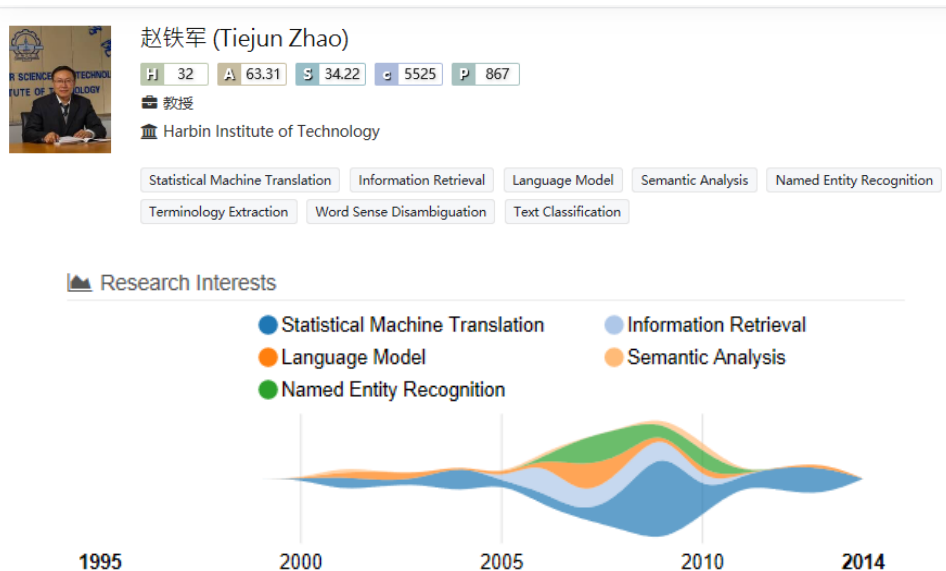
周明团队也为 Bing 搜索引擎提供了重要的技术支持，包括单词 breaker、情感分析、speller、解析器和 QnA 等 NLP 技术。他的团队创建了汉英、粤语的机器翻译引擎，为译者和 Skype 翻译。

最近，周明团队与微软产品团队紧密合作，在中国（小冰）、日本（Rinna）和美国（Tay）创建了知名的 chat-bot 产品，拥有 4000 万用户。他在顶级会议（包括 45+ACL 论文）和 NLP 期刊上发表并发表了 100 多篇论文，获得了 38 项国际专利。

我们根据 AMiner 大数据，筛选出周明发表论文中 citation 最高的几篇论文：

- Identifying synonyms among distributionally similar words SCOPUS WOS EI
Dekang Lin, Shaojun Zhao, Lijuan Qin, **Ming Zhou**
IJCAI (2003)
Cited by 208 Bibtex
- Improving query translation for cross-language information retrieval using statistical models EI SCOPUS WOS
Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, **Ming Zhou**, Changning Huang
SIGIR (2001)
Cited by 154 Bibtex <https://static.aminer.org/pdf/20170130/pdfs/index.txt>
- Chinese named entity identification using class-based language model EI
Jian Sun, Jianfeng Gao, Lei Zhang, **Ming Zhou**, Changning Huang
COLING (2002)
Cited by 125 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>
- Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations EI SCOPUS WOS
Jianfeng Gao, **Ming Zhou**, Jian-Yun Nie, Hongzhao He, Weijun Chen
SIGIR (2002)
Cited by 106 Bibtex <https://static.aminer.org/pdf/20170130/pdfs/index.txt>

● 赵铁军



赵铁军是哈尔滨工业大学计算机科学与技术学院语言技术研究中心教授，哈工大语言语音教育部-微软重点实验室副主任。

他的研究领域包括自然语言理解、机器翻译、基于内容的网页信息处理以及应用人工智能等。近年来主持完成国家自然科学基金、国家 863 项目、国防预研、省部委、国际合作项目等 20 余项。2009 年获得国防科技进步奖。近 4 年来获得软件著作权 7 项，在国内外刊物和会议上发表论文 150 余篇。

我们根据 AMiner 大数据，筛选出赵铁军发表论文中 citation 最高的几篇论文：

- Target-dependent Twitter Sentiment Classification SCOPUS WOS EI
Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, **Tiejun Zhao**
ACL (2011)
Cited by 623 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>
- HIT at TREC 2012 Microblog Track. EI
Zhongyuan Han, Xuwei Li, Muiyun Yang, Haoliang Qi, Sheng Li, **Tiejun Zhao**, Zhongyuan Han, Haoliang Qi
TREC (2012)
Cited by 38 Bibtex http://trec.nist.gov/pubs/trec21/papers/HIT_MTLAB_microblog_final.pdf

Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points

SCOPUS WOS EI

Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, Tiejun Zhao

COLING (2008)

Cited by 35 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points

SCOPUS WOS EI

Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, Tiejun Zhao

COLING (2008)

Cited by 35 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

● 刘洋



刘洋 (Yang Liu)

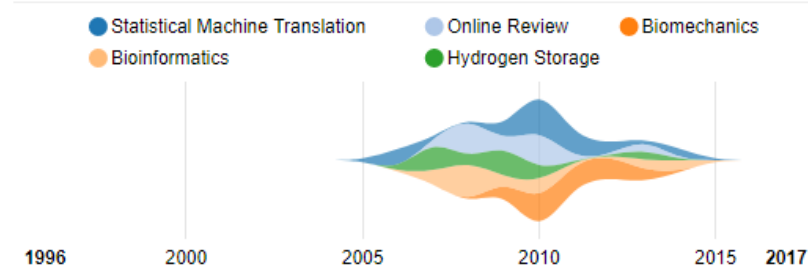
H 41 A 35.44 S 12.74 G 6608 P 505

Associate Professor

Department of Computer Science and Technology, Tsinghua University

Statistical Machine Translation Online Review Biomechanics Bioinformatics Learning Artificial Intelligence
Hydrogen Storage Machine Learning Finite Element Analysis

Research Interests



刘洋是清华大学计算机科学与技术系特别研究员。

主要研究方向是自然语言处理，近年来从事的科研工作集中在统计机器翻译领域。在自然语言处理和人工智能领域重要国际刊物 Computational Linguistics 和国际会议 ACL、EMNLP、IJCAI 和 AAAI 上发表 50 余篇论文，2010 年在自然语言处理领域国际顶级期刊计算语言学上发表国内第一篇长文，并在 ACL 上成为国内第一个做 tutorial 的学者。获 ACL 2017 杰出论文和 ACL 2006 优秀亚洲自然语言处理论文奖。承担 10 余项国家自然科学基金、国家重点研发计划、国家 863 计划、国家科技支撑计划和国际合作项目，2015 年获国家自然科学基金优秀青年项目资助。获得 2015 年国家科技进步二等奖、2014 年中国电子学会科学技术奖科技进步类一等奖等奖项。

我们根据 AMiner 大数据，筛选出刘洋发表论文中 citation 最高的几篇论文：

Tree-to-string alignment template for statistical machine translation

SCOPUS WOS EI

Yang Liu, Qun Liu, Shouxun Lin

ACL (2006)

Cited by 375 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

Log-linear models for word alignment

EI

Yang Liu, Qun Liu, Shouxun Lin

ACL (2005)

Cited by 125 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

Minimum Risk Training for Neural Machine Translation

EI

shiqi shen, yong cheng, zhongjun he, wei he, hua wu, maosong sun, yang liu,

ACL (2015)

Cited by 113 Bibtex <https://static.aminer.org/pdf/20170130/aclanthology/index.txt>

3.3 领域新星

AMiner 在机器翻译领域知名科研机构 and 实验室中挖掘了新生代的代表, 简单列举如下。

- **Maria Nădejde**

Maria Nădejde 是爱丁堡大学统计机器翻译 (SMT) 的 PhD, 师从 Philipp Koehn。主要研究兴趣包括利用文本的句法和语义结构来提高翻译质量。之前曾在 DFKI Saarbrücken 做过 HIWI, 在那里做了 MSc 论文, 提高了英语德语间翻译的质量。在 FBK Trento 的人类语言技术研究小组做过实习, 学习了使用基于类的语言模型来预测看不出的单词级别 n-grams。

- **Shuoyang Ding**

Shuoyang Ding 是约翰霍普金斯大学计算机科学系的 PhD, 师从 Philipp Koehn 和 Kevin Duh, 作为语言和语音处理中心的一员, 目前正在从事神经机器翻译的研究, 对自然语言处理领域中不同类别的问题感兴趣。在加入约翰霍普金斯大学之前, 获得了北京邮电大学的学士学位。在本科学习的最后一年, 在北京大学计算机科学与技术学院的语言计算与网络挖掘小组中与 Weiwei Sun 合作, 专注于语义分析和中文分词。

- **Adithya Renduchintala**

Adithya Renduchintala 是约翰霍普金斯大学语言和语言处理中心的 PhD, 师从 Philipp Koehn 和 Kevin Duh 主要研究领域是建立用户建模的计算模型, 特别是在语言学习方面。使用自然语言处理、机器翻译和机器学习技术来解决这个问题。此外, 还致力于为语言学习者构建语法校正模型。

Min Lin、Devansh Arpit、Jason Jo、Joseph Paul Cohen 等人都是 Yoshua Bengio 教授的 post-doc。

Vincent Dumoulin、Guillaume Alain、Bart van Merriënboer、Jessica Thompson、Julian Vlad Serban 等则是 Yoshua Bengio 教授的 PhD。

- **Brian DuSell**

Brian DuSell 是圣母大学计算机科学专业的 PhD, 在 David Chiang 的 NLP 研究小组中学习自然语言处理。主要兴趣是一般解析和机器翻译。目前正在调查使用基于堆栈的 RNN 架构, 或“神经推下自动机”, 以提高对人类语言的句法分析。

- **Kenton Murray**

Kenton Murray 是圣母大学计算机科学与工程项目二年级 PhD, 师从 David Chiang。主要对机器学习和自然语言处理的统计方法比较感兴趣, 特别是机器翻译。主要致力于阿拉伯语的翻译, 将非结构化文本合并到事件检测中。

- **张飏**

张飏师从厦门大学苏劲松教授, 并在联合课题项目中受苏州大学熊德意教授的指导, 所

属实验室长期致力于人工智能尤其是自然语言处理方向的研究,张飏本人则长期致力于机器翻译领域的研究,借助深度学习的方法来建模源句子的语义表示以及不同语言对之间的翻译关系。读研期间,张飏作为骨干研究员参与多项国家及省自然科学基金项目,并在 AAAI、IJCAI、EMNLP、COLING、TASLP 和 INS 等人工智能和自然语言处理顶级会议和期刊上发表学术论文。

● 沈世奇

沈世奇目前在腾讯工作。博士毕业于清华大学自然语言处理组,由刘洋教授和孙茂松教授共同指导,主要研究兴趣的机器翻译和结合深度学习的自然语言处理,并参与了国家 863 项目中的相关工作。

4

application

应用篇



4 应用篇

机器翻译技术较早的被广泛应用在计算机辅助翻译软件上,更好地辅助专业翻译人员提升翻译效率,近几年机器翻译研究发展更为迅速,尤其是随着大数据和云计算技术的快速发展,机器翻译已经走进人们的日常生活,在很多特定领域为满足各种社会需求发挥了重要作用。

提到机器翻译绕不开的一个话题是谷歌翻译。谷歌翻译是谷歌公司推出的针对文本、语音、图像以及实时视频的多语种翻译服务。该项目始于 2001 年,上线初期采用与其他同类型公司(例如雅虎)类似的机器翻译系统,但是翻译精度并不理想。2004 年下半年起,随着 Franz Josef Och 成为其首席科学家,谷歌翻译进入迅速发展阶段。在 2005 年的 NIST 机器翻译系统比赛中,谷歌翻译一举拿到第一名。在 2006 年的比赛中,谷歌翻译几乎包揽全部比赛项目的第一名。根据维基百科公布的数据,谷歌翻译支持 103 种语言,每天为过两亿人提供免费的多种语言翻译服务,为业界的标杆。

近两年来,谷歌机器翻译的突破让人目不暇接。2016 年 9 月谷歌发布 Google 神经网络机器翻译系统,简称 GNMT,使用当前最先进的训练技术,能够实现迄今为止机器翻译质量的最大提升。2017 年 6 月再次宣布又在机器翻译上更进了一步,实现了完全基于 Attention 的 Transformer 机器翻译网络架构,并且还在 WMT 2014 的多种语言对的翻译任务上超越了之前 Facebook 的成绩,实现了新的最佳水平。北京时间 2017 年 3 月 29 日,经过大幅优化了中国用户体验的 Google 翻译正式宣布在中国上线,在中国不用“翻墙”就可下载使用,实景翻译、语音翻译、离线翻译,和点按翻译(仅限 Android 平台)等功能赫然在目。

下面我们按照媒介来对机器翻译的应用进行一个简单分类。

● 文本翻译

机器文本翻译顾名思义就是将源语言文字转换为目标语言文字,即书面语及其翻译,在目前的机器翻译中应用最为广泛。

目前,文本翻译最为主流的工作方式依然是以传统的统计机器翻译和神经网络翻译为主。Google、Microsoft 与国内的百度、有道等公司都为用户提供了免费的在线多语言翻译系统。将源语言文字输入其软件中,便可迅速翻译为目标语言文字。Google 主要关注以英语为中心的多语言翻译,百度则关注以英语和汉语为中心的多语言翻译。另外,即时通讯工具如 GoogleTalk, Facebook 等也都提供了即时翻译服务。

速度快、成本低是文本翻译的主要特点,而且应用广泛,不同行业都可以采用相应的专业翻译。但是,这一翻译过程是机械的和僵硬的,在翻译过程中会出现很多语义语境上的问题,仍然需要人工翻译来进行补充。

● 语音翻译

语音翻译可能是目前机器翻译中比较富有创新意思的领域,吸引了众多资金和公众的注意力。亚马逊的 Alexa、苹果的 Siri、微软的 Cortana 等,我们越来越多的通过语音与计算机进行交互。

正确的从一种语言的源语音翻译成不同的目标语言需要经过四步。首先利用语音识别,

将音频转换为文本；其次利用算法将口语优化为更标准的文本，使它更适合机器翻译，然后通过文本翻译的方法进行翻译；最后文本到语音转换，必要时输出译文的音频。

应用比较好的如语音同传技术。同声传译广泛应用于国际会议等多语言交流的场景，但是人工同传受限于记忆、听说速度、费用偏高等因素门槛较高，搜狗推出的机器同传技术主要在会议场景出现，演讲者的语音实时转换成文本，并且进行同步翻译，低延迟显示翻译结果，希望能够取代人工同传，实现不同语言人们低成本的有效交流。

科大讯飞、百度等公司在语音翻译方面也有很多探索。如科大讯飞推出的“讯飞语音翻译”系列产品，以及与新疆大学联合研发的世界首款维汉机器翻译软件，可以准确识别维吾尔语和汉语，实现双语即时互译等功能。

● 图像翻译

图像翻译也有不小的进展。谷歌、微软、Facebook 和百度均拥有能够让用户搜索或者自动整理没有识别标签的照片的技术。图像翻译技术的进步远不局限于社交类应用。医疗创业公司可以利用计算机阅览 X 光照片、MRI（核磁共振成像）和 CT（电脑断层扫描）照片，阅览的速度和准确度都将超过放射科医师。而且更图像翻译技术对于机器人、无人机以及无人驾驶汽车的改进至关重要，福特、特斯拉、Uber、百度和谷歌均已在上路测试无人驾驶汽车的原型。

除此之外还有视频翻译和 VR 翻译也在逐渐应用中，但是目前的应用还不太成熟。

5 trend

趋势篇



5 趋势篇

从全局热度看，machine translation、statistical machine translation、natural language processing、language model 等是整体关注的热点。

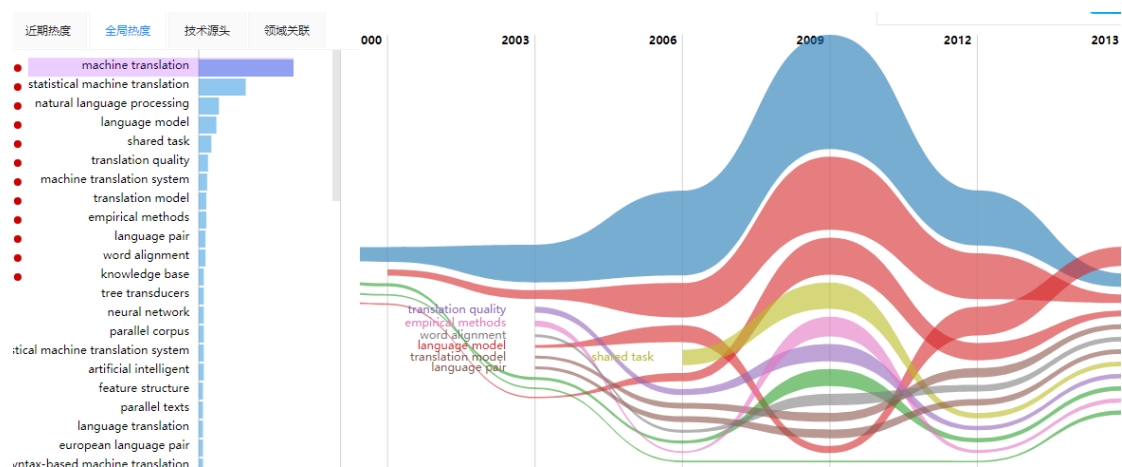


图 16 机器翻译领域全局热度

近期的关注热点则是集中在 language model、machine translation、new story 等领域。

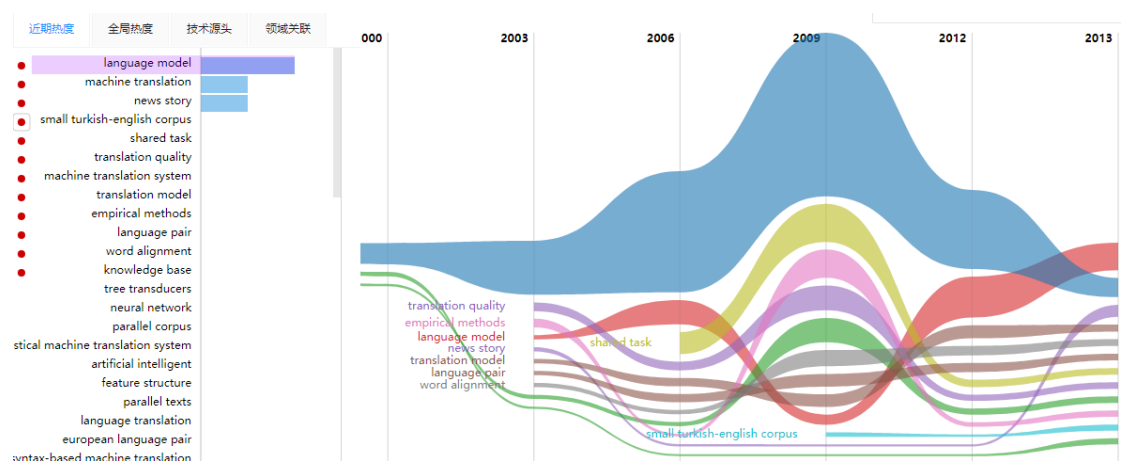


图 17 机器翻译领域近期热度

随着全球化和互联网迅速发展，跨语言的网络资源不断呈几何级数增长，迅速改变着信息传播的方式，极大地刺激了全球机器翻译产业的发展。随着产品技术的进步，人们对机器翻译的可接受度在迅速提高，同时也越来越清楚机器翻译能够做什么，以及应当怎么去做才能最大发挥机器翻译效能，总得来说，机器翻译的发展将呈现以下趋势。

● 实用化

机器翻译逐渐实用化，被应用到生活场景中。过去机器翻译更像是一个“智能的词典”，帮助人们阅读外文网页内容。现在随着语音和图像识别技术的进步，机器翻译可以更多地与生活场景结合。比如人们出国时可用百度翻译了解菜单、店名、商品信息，看美剧时可以用电脑进行字幕翻译，通过拍照直接翻译出一朵花的名字，再比如利用机器人进行多语翻译采访等。

● 多模态

机器翻译的多模态融合。多信息融合是一个重要的研究趋势，机器翻译是一个热门的研究领域，随着训练数据规模的增加，各种 NN 模型的效果也取得了突破的进展，Google 和百度均已部署上线 NMT 系统，融合图像、音频、视频、文本等各种模态数据的机器翻译会是将来的一个发展方向。

● 多语言

机器翻译将从资源丰富语言向匮乏语言迁移。以往的机器翻译主要集中在英语、汉语等应用范围较广、使用人数较多的语种上，对于使用频率不是很高的小语种缺乏关注度。近几年来，机器翻译的语言不断扩大，语种逐渐向小语种开始倾斜。Google 于 2011 年 1 月正式在其 Android 系统上推出了升级版的机器翻译服务，支持近 20 种的多国语言之间的互译，微软的 Skype 于 2014 年 12 月宣布推出实时机器翻译的预览版、支持英语和西班牙语的实时翻译，并宣布支持 40 多种语言的文本实时翻译功能，国内百度翻译、科大讯飞等也在加大多语言翻译系统的开发。

● 网页端向移动端转移

机器翻译由网页端向移动端转变。早期的机器翻译主要是在网页上进行，但是随着计算机技术的发展和智能手机的普及，移动端的应用越来越方便，主要的公司诸如 Google 有道等公司纷纷推出了自己的 APP 等产品。

● 垂直领域结合更紧密

垂直场景对于机器翻译的需求更甚，从出国旅行，到国际文化交流，再到对外贸易，语言障碍是一个天然痛点，目前很多翻译类的产品将机器翻译和 OCR 技术以及语音识别技术进行结合，可以实时的通过摄像头来翻译外文指示牌、菜单、说明书等，也可以结合语音技术进行对话翻译，从而实现不同语种的无障碍交流，与各行各业实现更为紧密的结合将会是机器翻译的一个方向。

参考文献

- [1] 刘群.统计机器翻译综述.[J].中文信息学报第 17 卷第 4 期.2003
- [2] 刘群.机器翻译技术现状与展望.[J].集成技术第 1 卷第 1 期.2012
- [3] 冯志伟.机器翻译研究.[M].北京：中国对外翻译出版社.2004
- [4] 刘洋. 基于深度学习的机器翻译研究进展[J].中国人工智能学会通讯
- [5] 杜金华，张萌，宗成庆，孙乐.中国机器翻译研究的机遇与挑战—第八届全国机器翻译研讨会总结与展望.[J].中文信息学报第 27 卷第 4 期.2013
- [6] Bond, *Francis.Toward a Science of Machine Translation.*[EB/OL].2014
- [7] Chomsky, *Andrew Syntax.Agenerative Introduction.*[M].USA：Blackwell Publishing.2014

版权声明

AMiner 研究报告版权为 AMiner 团队独家所有，拥有唯一著作权。AMiner 咨询产品是 AMiner 团队的研究与统计成果，其性质是供用户内部参考的资料。

AMiner 研究报告提供给订阅用户使用，仅限于用户内部使用。未获得 AMiner 团队授权，任何人和单位不得以任何方式在任何媒体上（包括互联网）公开发布、复制，且不得以任何方式将研究报告的内容提供给其他单位或个人使用。如引用、刊发，需注明出处为“AMiner.org”，且不得对本报告进行有悖原意的删节与修改。

AMiner 研究报告是基于 AMiner 团队及其研究员认可的研究资料，所有资料源自 AMiner 后台程序对大数据的自动分析得到，本研究报告仅作为参考，AMiner 团队不保证所分析得到的准确性和完整性，也不承担任何投资者因使用本产品与服务而产生的任何责任。