哈工大自然语言处理研究进展

李 生

哈尔滨工业大学 哈尔滨 150001

摘要：本文阐述了自然语言处理研究的重要意义，介绍了哈尔滨工业大学在自然语言处理领域的研究历史和研究现状，并综述了哈工大在该领域各方向的研究进展。

The Progress on Researches of Natural Language Processing in Harbin Institute of Technology

Li Sheng

Harbin Institute of Technology Harbin 150001

ABSTRACT: This paper focuses on the importance of natural language processing (NLP) first. We present a review on the research stages and the state-of-the-art about NLP researches in Harbin Institute of Technology (HIT). The field progress of each branch in the researches is also summarized.

1. 自然语言处理对计算机学科发展的贡献

用计算机自动处理语言，是一个伴随着计算机而诞生的孪生兄弟。从“行编辑”到“Word-star”再到现在的 Office，自然语言处理技术一直是推动着计算机应用不断普及和深入的一个重要推动力。事实上，自然语言处理研究对于计算机学科发展的重要性远远不止于此。

首先，从理论上看，语言是思维的外壳。自然语言的自动处理研究正是计算机学者对于人类智能的探索。在这一探索过程中，计算机研究者逐渐认识到自然语言的处理是人工智能中最具挑战性的课题。

目前的自然语言处理，已经从初期的文字处理发展到语音识别、语音合成、OCR 识别、句法分析、自动文摘、问答系统、信息检索、机器翻译等多个研究分支。所使用的技术也从初期的产生式系统发展到统计模型、机器学习等方法。自然语言处理的研究成果不仅正在服务于各种应用，而且还促进了如生物信息学等一些新兴学科的发展。

同时，对于自然语言处理的追求并不仅仅是发展了计算机中的人工智能（或者计算语言学）某个单一学科。对于自然语言处理的认识，正促使计算机的体系结构发生着变化：下一代的计算机已经把能够处理自然语言的作为其中的一个追求目标。

从体系结构到操作系统，从互联网到数据库，计算机在经历了20世纪的迅速发展之后，所要处理的对象已经从处理传统的数据提高到复杂信息内容的处理。因此，刚刚进入21世纪不久，我们发现自然语言的自动处理技术已经成为计算机科学必须直面的下一个难题。据美国计算语言学学会(ACL)的不完全统计，全世界目前已有110所大学共开设了173门自然语言处理类别的课程，课程种类接近100种。全美的计算机学科的前10所大学中，自然语言处理研究都是其中的一个主要方向。而在微软、IBM、贝尔实验室等全世界著名的公司和研究机构中，都有为数众多的自然语言处理研究人员。另据报道，美国国防部高级研究计划局(DARPA)所支持的TIDES (Translingual Information Detection, Extraction and Summarization)项目仅在2001-2002度的投资就超过1亿5千万美元，而DARPA同时资助的类似的同等规模自然语言处理项目还有若干个。所以，从实践上看，自然语言处理的研究已经成为计算机学科发展中的另一个主要增长点。

正是人们看到了自然语言处理在未来信息社会中所扮演的重要角色，所以从政府机构到各大公司都投入了大量资金进行研究开发，并作为长期项目来管理。毫无疑问，从市场价值到社会意义，自然语言处理技术都将会充分展示其潜力，人们的生活将会随着自然语言处理技术的应用和改进而改变。

2. 哈工大自然语言处理研究的历史和现状

哈工大自然语言处理研究事业萌芽于二十世纪五十年代，重整于七十年代末，发展于八十年代，壮大于九十年代。今天，在回顾哈工大自然语言处理研究历史的时候，不能不提及那些为开拓我们研究领域做出了重要贡献的前辈。

早在五十年代，全国集中人力在北京中科院计算所开展俄汉机器翻译研究，就有我校俄语教研室王畛教授参加。1979年，王开铸教授从美国考察归来，与哈工大俄语教研室合作，率先展开了俄汉题录机器翻译研究，并于1981年推出HIT-80俄汉题录翻译系统。尽管当时所用的计算机没有汉字系统，但课题组克服了很多困难，终于取得成功。当时受邀请前来参观的专家对该系统给予了很高评价。

王开铸教授从1985年开始研究汉语理解，前后涉猎了从汉语自动分词到中文自动文摘等多个研究方向，取得了许多优秀成果。在863计划的支持下，1988年研制成功了一个基于理解的固定段落中文问答实验系统CQAESI。后来又与清华大学合作，研制成功了基于固定字、词域的段落理解问答系统。1987年开始研究汉语分词问题，并将分词理论与语音识别相结合，在1990年研制成功863-四达语音识别系统，在由国家科委高新司主持的鉴定会上被评为全国领先水平。从九十年代开始，王开铸教授研制了中文自动文摘系统，在国家863计划的支持下，先后实现了基于理解的中文文摘系统和基于统计的任意文本文摘系统，通过了863专家组主持的鉴定，评为全国领先水平。此外，王老师还受上海《新民晚报》的委托，研制成功了中文校对系统。在该系统中应用了自然语言理解技术、自动分词技术，提出了碎片思想用于中文文本查错。在交付用户的测试过程中，该系统查出了经过三较、已经发表的一篇文章中的一个错别字，折服了挑剔的用户。

舒文豪教授，1979年至1981年在美国从师于国际结构模式识别创始人傅京孙教授，专

门对汉字识别进行研究。1982 年回国后，开展了模式识别方向的研究，领导创建了文字识别研究室，使哈工大成为国内开展该方向研究最早的几个单位之一。其中手写体汉字识别的研究是其中一个主要方向。在研究中，既有软件研究开发，也有硬件的研究与设计，其研究成果一直处于国内先进的行列，1985 年全国第一届计算机应用成果展览会上就展出了哈工大文字识别的研究成果，引起了轰动。随后本方向研究被列入国家“七五”攻关项目，经过不断努力、攻关，1992 年“手写汉字图文输入装置”通过机电部组织的鉴定，成为我国第一个从硬件到软件都立足于国内器材的产品。1996 年研制出国内第一台具有笔输入功能的 PDA。舒老师在科研方面取得突出成果，也培养出很多优秀人才。

李生教授从 1985 年开始汉英机器翻译研究，是我国最早研究和开发汉英机器翻译系统的学者之一。在李生教授的领导下，完成了汉英机器翻译系统 CEMT-I，该系统于 1989 年 5 月鉴定，成为我国第一个通过技术鉴定的汉英机器翻译系统，获部级科技进步二等奖。随后，李生教授带领课题组成员以合作方式先后完成了汉英机译系统 CEMT-II、CEMT-III 系统，后者于 1993 年 5 月在北京鉴定，受到与会专家的好评，并获部级科技进步二等奖。1993 年秋季原航天工业总公司投资开发“达雅翻译工作站（机器辅助翻译系统）”。这是哈工大在自动翻译软件商品化方面进行的一次可喜尝试。该软件先后连续三年在北京计算机产品交易会上展出，获得好评，并获 1997 年部级科技进步二等奖。从 1994 年起，李生教授与课题组其他成员一起在国家 863 高技术研究发展计划的支持下，开展了汉英—英汉双向机器翻译研究，先后实现了 BT863-I 和 BT863-II 双向机译系统，通过鉴定并获奖。随着研究的不断深入，课题组在机器翻译相关的一系列技术方面进行了广泛探索，包括汉语自动分词与词性标注、汉语和英语句法分析、词义消歧、英汉双语语料库加工等。

哈工大语音处理研究开始于八十年代初，徐近霈教授领导的研究组完成了汉语孤立词电话号码查询演示系统，当时香港大公报曾做了相关的报道；八十年代中期，赵国田教授从日本留学归来，在国内较早地开展汉语大词表语音识别的研究，取得了较大的成绩，当时的李鹏总理曾题词表示祝贺。八十年代后期开始，哈工大在汉语单音节识别、噪声下的语音识别等方面开展了卓有成效的工作；朱志莹、王承发教授等研究了声控篮球比赛临场统计系统，研制了抗噪声话筒，使语音识别技术直接应用于实际比赛，1991 年该成果获得部级科技进步二等奖；徐近霈教授等完成了汉语文本读入系统，1993 年获部级科技进步二等奖；从九十年代中期开始，王承发教授等重点开展了高噪声下语音识别的研究，完成了国内首创的高噪声背景下命令语音识别系统，1995 年获级科技进步三等奖。

现在哈工大在自然语言处理领域拥有以王晓龙教授为代表的一批年轻的优秀专家和人才，在中文信息处理、机器翻译、语言内容管理与信息检索、模式识别、语音处理等各个方向上都进行着广泛而深入的研究。哈工大计算机学院在自然语言处理研究方面共有研究人员 26 名，其中教授 7 名（博士生导师 6 名）、副教授 9 名，是我国该领域研究的一支重要力量。在各个研究分支中，共有在读博士生 40 多名，硕士生 60 多名。

目前，哈工大自然语言处理研究主要分布在 5 个研究室当中，这些研究室分别或者联合承担了智能化中文信息处理、汉英-英汉双向机器翻译、语言内容管理与信息检索、文字与人体生理特征识别、语音识别和语音信息处理等方向上的一系列科研课题。我们不断探索多学科交叉的新领域，通过联合攻关，努力在重大科研项目上有所突破，为我国计算机科学特别是自然语言处理领域的发展做出应有的贡献。

3. 计算机学院自然语言处理各研究方向进展

3．1 机器翻译

哈工大机器翻译研究始于 1985 年，在原航天工业部和国家 863 计划支持下，对汉英机器翻译技术和汉英-英汉翻译技术进行了深入的探索，先后研制了多个自动翻译系统，多次获得部级科技进步奖。

2000 年以来，机器翻译研究室在坚持原有学术探索的同时，更加重视机器翻译系统的实用化工作：即如何改进利用现有技术为应用服务。而这一工作的核心就是基于统计的自然语言学习技术。

针对机器翻译的特点，我们首先进行的是基于汉英平行语料库的双语知识获取。这项研究引起了当时的微软中国研究院的兴趣，成为了 2000 年 6 月成立的微软-哈工大机器翻译联合实验室的第一个合作研究计划。在这一框架下，我们完成了汉英双语语料库的多级对齐加工技术（包括句子对齐、词对齐和结构对齐），基于对齐的英汉翻译知识获取技术，完成了 6 万句词汇对齐的汉英双语语料库建设，其中 2 篇论文在 COLING2002 上宣读。

利用这些研究成果，我们构建了汉英双向机器翻译 MTS2000 系统。这一系统虽然沿用了经典的“分析-转换-生成”策略，但是由于设计上采用模块化的思想，所以系统集成了我们多年的研究成果：包括汉语分词、英语词法分析、词性标注、句法分析、词义消歧、翻译知识获取等等，因而也就成为我们的一个机器翻译研究演示平台，也成为了研究成果的测试比较平台。

从 2001 年下半年开始，在国家 863 计划的统一协调下，我们参与了面向奥运的多语机器翻译系统的研制。经过仔细调研，根据相关单位的意见和专家建议，将课题重点放在面向体育新闻的多语汉外机器翻译系统，初期完成了面向田径和球类的体育新闻的汉英机器翻译原型系统。在将现有翻译技术服务于奥运这一过程中，我们进一步实现了以下技术：

（1）面向特定领域的基于模板的机器翻译技术；设计并实现了以动词为核心、基于语义的信息翻译模版，目前规模包括体育领域常用 336 个动词的 837 个模版。

（2）基于汉语分析的特定信息的获取和翻译技术；采用级连的 Markov 模型实现了汉语的浅层句法分析，采用统计和规则相结合的方式实现了汉语句子主干的识别，初步解决了体育新闻中的关键信息识别问题。

在研究中，我们感到目前制约机器翻译系统在奥运中的应用瓶颈包括两个方面：一个是汉语分析技术，另一个是翻译系统的快速移植。为此，机器翻译研究室从 2002 年底开始了新一轮两个方面的攻关。

我们首先加强了有关句法分析研究。英语句法分析方面，我们通过微软研究院的帮助，实现了目前英语最好的句法分析器。在汉语句法分析方面，我们首先进行了 2 万汉语句树库的建设，以解决统计训练的样本。从目前实验结果来看，在相同的条件下，基于最大熵的模型在汉语句法分析中性能较优。

对于翻译系统的快速移植，我们从技术上探索了基于实例的机器翻译技术。利用已有的

汉英双语语料加工技术，我们快速完成了基于实例的英汉机器翻译原型系统。为了解决机器翻译系统所需的翻译实例规模，我们在863的支持下在国家863计划的支持下开始构建面向旅游、体育、商务领域的汉/英/日三语平行语料库。语料库总规模为200万字（词），预计在2004年下半年完成，现在已经完成了大约80万字的内容。

3．2 中文信息处理

哈工大自然语言处理研究室在中文信息处理研究方面取得了以下重要进展：

（1）拼音智能输入

哈工大拼音智能输入的研究是从1984年开始的，1990年王晓龙教授首先提出语句输入的思想，研究了语句输入的总体设计、人机界面、局内编辑、机器学习等有关内容，并完成了第一个拼音语句输入系统和语句级语音输入系统，当时中文信息学会理事长陈力为院士给予了充分的肯定。该技术在国家863鉴定为“国际领先地位”，并先后在1995年、1996年、2002年授权给美国微软、日本佳能泰克、日本富士通等公司。目前语句输入的方式已经成为键盘输入的主流技术。

（2）自动文摘技术

哈工大早在80年代开始这方面的研究工作，包括中文文摘和英文文摘两方面的研究，目前我们学院实现的文摘系统主要有两个：一个是基于多知识源融合汉语自动文摘系统InsunAbs，主要采用多知识源融合的方法，将文档中所包含的多种特征，包括文章的统计特征、结构特征、语法特征、语义特征等，按特定的人工智能方法进行了集成，同时加入了文档自动分类、人名识别等模块，克服了当前文摘系统算法复杂度高以及生成的文摘可读性差的问题，在保证系统实时性能的同时，有效地提高系统的准确性以及所生成摘要的流畅性与内在逻辑连贯性。同时通过建立一个文摘系统性能自动定量评测体系，实现了文摘系统性能的自动优化，从而保证了文摘系统的可移植性。该系统在国家网络安全项目验收时评为优。另一个系统是由刘挺博士主持开发的汇总多篇相同主题的文章的多文档文摘系统。

（3）智能化中文信息处理平台

该项目为国家863重点项目，集合了哈工大本研究领域的自然语言处理课题组、机器翻译课题组、模式识别（文字识别）课题组、语音识别课题组、计算机新技术研发中心5个课题组的力量，集成了各课题组的多语种机器翻译、语音识别与合成、文字识别技术、语言处理技术、基于内容的Internet信息搜索、处理和理解技术、大型基础资源库等。在研究中，我们以多年研究的技术和资源为基础，进一步完善、融合了这些关键技术并在某些方面取得突破性进展，从而研制出面向网络环境的新型智能化中文信息处理平台。其技术特点包括：网络环境—强调海量和动态；开放式框架—不拘泥于一家技术，融合多方优势，避免重复劳动，取长补短，强强合作；系统可定制—为各类通用、专用用户构造适合他们需求的资源库和关键技术。该项目的实现做到了资源开放（如语料库共享）、接口开放（如统一标准）、关键技术服务开放（如功能借用）。

（4）基于内容的智能检索技术与自然语言问答系统

为了满足人们以自然的方式快速、准确地在互联网上获取信息这一强大需求，针对当前

信息检索技术存在的主要问题，考察了国际信息检索技术的发展动态，提出了研发以自然语言为接口、从网络在线文档中自动抽取答案并返回用户的新一代智能化信息检索系统的设想。在研究开发过程中，已经实现了具备接受用户的自然语句查询直接返回相关问题答案的功能，具备了问答系统的形式，为正在进行中的国家863项目的智能搜索引擎部分的研发工作奠定了基础。最终目标是完成一个实用化的面向旅游和体育领域的问答系统。

3．3 信息检索和语言内容管理

信息检索研究室在语言内容管理的一系列关键技术方面做了大量探索。这些探索研究包括：

（1）语言信息的搜集和校对，如汉英句对检索，目前已经积累了40万中英文双语句对，并开发了慧通英文辅助写作系统产品。中文自动校对系统作为863课题的子课题，其实现的性能与MS Word相当。

（2）语言内容分析，如语言内容的标注和语义分析，进行了汉语依存关系分析和基于HowNet的汉语语义标注。在依存关系分析研究中，采取了汉语依存骨架分析，开发了5万句汉语依存骨架树库。

（3）语言内容的索引和检索，触摸手写作为输入手段，实现了文本检索系统，并在校园e点通中得到实际应用。此外还实现了网上热点话题自动发现系统。

（4）语言内容的分类、过滤和推送，研究实现了垃圾短信模糊过滤系统、中英文文本分类系统、网上有害信息过滤系统等。

（5）语言内容的聚类和去重，研究实现的网页去重算法具有线性时间复杂度，准确率达到99%。

（6）问答系统，受国家自然科学基金资助的资助，目前正在开展开放域问答系统的研究，特别注重从自然语言文本中获取知识，以及对这些知识进行加工推理基础性的研究工作。

3．4 文字识别

文字识别研究室的研究起步早，1980年起就成为当时少数几个从事文字识别的研究工作的单位之一；研究面宽，既开展方法研究，也研制相关设备，在软件硬件开发方面均具有一定实力，在各个时期都有相应的研究成果。其代表性成果包括：1992年率先研制出一个从硬件到软件都立足于国内器材的联机手写体汉字输入分析系统。1996年研制出我国第一台具有笔输入功能的掌上型电脑。1999年研制出袖珍型扫描史英汉词典。曾获国家发明专利一项、国家科技进步三等奖一项、部级科技进步一等奖一项、部级科技进步二等奖四项、部级科技进步三等奖二项。历年来在国内外发表有关汉字识别研究论文200余篇、专著一部。研究室现有研究人员9人，其中教授3人、副教授2人、讲师2人、助教2人。

目前研究室与自然语言处理相关的研究集中在字符识别的理论与实践方面。在理论上，研究统计学习理论、自适应识别理论和多识别器融合理论，在现有汉字识别系统的基础上实现自由草书书写的识别，以及系统的在线自适应识别。在应用方面，开展多方面的字符/符号识别

的应用研究：在单字符识别的基础上，研究西文字符识别技术，以及字符识别模型的级联技术，将字符的识别、分割和语言模型处理统一在一个模型实现，完成了汉字/西文字符自由整句输入识别系统，提高识别输入系统的性能。同时开拓联机字符识别的研究领域，利用字符识别技术研制手写公式识别系统，进行手写编辑手势识别、手画图形识别等多种识别技术的研究；此外，还研究了固定图像和移动图像中的汉字/西文字符/数字的识别技术。

3．5 语音处理

哈工大语音处理研究室成立于二十世纪八十年代初，是国内较早开展此方面研究的单位之一，由3个课题组逐步合并演变而成。目前有教师4名，其中博士生导师2名，教授1名，讲师1名。在读博士研究生8人，硕士研究生10余人。国家计算机信息内容安全重点实验室语音监控研究室设在本研究室。先后完成了国家攻关项目、国家自然科学基金项目、国家863计划项目等，获省部级科技进步奖5项。发表论文100余篇，出版专著2部。目前承担2项国家自然科学基金项目，1项教育部跨世纪人才培养计划项目。研究室拥有微机近20台、良好的网络资源及语音处理方面相关的软件、图书资料等。

语音研究室在语音识别与处理方面开展了如下方面的研究：

（1）Robust 语音识别研究

环境噪声及说话人心理和生理情况的变化严重影响了语音识别系统的实际应用。本研究室从九十年代初开始，在国内率先开展高噪声环境下的语音识别。目前在国家自然科学基金等项目资助下，重点开展噪声下的语音识别、变异语音识别、情感语音合成等方面的研究。本研究在军事指挥、生产现场、单兵数字系统和HPC的语音控制等方面具有广阔的应用前景。

（2）语音信息安全

近年来，因特网的应用变得越来越广泛，随之而来的信息安全问题也日益突出，并逐渐成为社会性问题。如何对其中的语音内容进行有效地控制，对国家的信息安全建设具有重要的意义，本研究室目前在国家有关项目的资助下重点开展此方面的研究。

此外，研究室目前正着手开展音频信息检索技术、基于语音识别技术实现发音矫正，指导发音学习的技术。

4 展望

哈尔滨工业大学计算机学院在自然语言处理方面人才济济、实力雄厚，在该领域的各个分支上都开展了全面而深入的研究，许多研究成果已经实用化。应该说，哈工大自然语言处理研究者为我国计算语言学领域的发展做出了自己应有的贡献。当前，随着互联网的普及和发展、信息科学与其他科学技术领域的相互渗透和交叉，自然语言处理研究正在面临着前所未有的机遇和挑战。面对新形势，哈工大自然语言处理研究者正在密切跟踪国内外相关学科的动态，做好准备，加强合作，为共同承担重大科研课题的研究而努力。

纵观近年来国内外自然语言处理研究的发展，一些新的潮流和动向正在出现。就文本处理方面来说，我们看到：

（1）自然语言处理的研究正在从解决信息的输入输出问题，如汉字识别、汉字编码、拼音输入、激光照排输出等，向内容管理发展。随着Internet上可利用的自然语言资源的增多，如何对这些海量语言资源进行分类、检索、摘要发布等，使之能够更方便快捷地被人们更多更好地利用，这个问题越来越突出。

（2）语言处理的深度不断加深，从字到词，到句，到篇章大语境的处理，从词法、句法到词义、句义，乃至到语言所负载的知识的处理，越来越深。正在从语言表层处理向理解和知识挖掘发展。

（3）理解的路线和统计的路线在走向融合，最早的人工智能方法试图探索语言深层含义，但是由于条件所限，当时处理的数据规模小，不够真实。后来的统计方法很多运行在语言的表层，但是处理的数据量大，语料真实。现在技术的发展有向人工智能回归的趋势，将统计的方法运行在语言的深层结构上，在海量的矛盾的真实数据环境下进行语言理解的探索。

（4）研究和评测的资源正趋于标准化和共享。大规模语料库和统计方法的广泛应用改变了自然语言处理的传统方式，使用大规模多层次加工的语料库作为统计模型的训练和测试样本已经成为一种新的工作模式。语料库建设的重要意义已经越来越被研究者所认识，但是目前国内研究还需要建立共享的、开放的、尽可能规范的基础数据集合，使我国的研究资源设置更趋合理，通过统一评测、共享和交流不断提高我们的研究水平。

在语音研究领域，目前语音识别从总体上看，基础研究以语音产生的模型、不利环境下的Robust语音识别等为主；应用研究以限定领域的应用为主。出现了一些新的研究热点和方向：多通道信息融合，语音内容深层处理如音频检索、记录、分类、自动文摘、标题等，语种识别，嵌入式及资源有限时的语音识别，网络环境下的语音识别和话者证实，基于语音的情感处理以及语音的发音矫正训练等。

哈尔滨工业大学计算机学院自然语言处理研究相关主页连接：

机器翻译：http://mtlab.hit.edu.cn

信息检索：http://ir.hit.edu.cn

语音处理：http://splab.hit.edu.cn