第37卷 第5期
2013年9月

燕山大学学报
Journal of Yanshan University

Vol. 37 No. 5
Sept. 2013

文章编号：1007-791X (2013) 05-0377-08

自然语言处理的研究与发展

李　生*

（哈尔滨工业大学 计算机科学与技术学院，黑龙江 哈尔滨 150001）

摘　要：自然语言处理是信息技术最重要的研究方向之一，汉语自然语言处理应该包括对字、词、句子以及段落与篇章的处理。文中阐述了自然语言处理的研究方法及其现状，特别指出了汉语处理存在的特殊问题。概述了自然语言处理的主要研究方向，总结了我国在自然语言处理方面已经取得的成就与存在的问题，提出了对自然语言处理研究未来工作的思考。

关键词：自然语言处理；中文信息；共性技术；展望

中图分类号：TP391　　**文献标识码：**A　　DOI：10.3969/j.issn.1007-791X.2013.05.001

0　引言

信息同能源、材料一起构成经济发展与社会进步的三大战略资源。信息技术正在推动和改变人类的生产、生活甚至是思维方式。信息是无形的，但它可以用语言来表达。语言是信息的载体，语言是文化的支柱，语言是人类思维、沟通与交流的工具。语言技能是一种人力资本。语言与经济、文化、教育，与社会发展和人类进步有着紧密的关系。

自然语言处理是信息技术最重要的研究方向之一。中文信息处理的研究内容是利用计算机对中文的音、形、义等语言文字信息进行加工和操作，包括对字、词、短语、句子、篇章进行输入、输出、识别、转换、压缩、存储、检索、分析、理解和生成等。它是语言学、计算机科学、认知科学、数学等多学科交叉的边缘学科。

自然语言处理是计算机应用的一个分支，是人工智能的一部分，但计算机技术和人工智能技术又都从属于信息技术。

自然语言处理通常是指用计算机对人类的自然语言进行有意义的分析与操作。

对于自然语言来说，有意义的最小单位应该是词，而对汉语来讲基本单元是字。对于汉语自然语言处理应该包括对字、词、句子以及段落与篇章的处理，见图1。

图1　自然语言处理分解示意图

Fig. 1　Decomposition diagram of natural language processing

自然语言处理的研究方法分成基于规则和基于统计的两种，基于规则的方法是人工获取语言规则，而基于统计的方法则是通过对大规模语料库的统计分析，实现对自然语言的处理。

处理的过程通常应该是：

语言问题—形式化表示（模型）—算法转换—程序编制—机器运行—结果输出—系统评测

中文（汉语）信息处理所遇到的特殊问题是：

1) 句子中单词的切分；

2) 时态、语态、语气等没有严格的形式标记；

3) 句子成分的省缺及指示代词的频繁出现；

4) 语言资源的缺乏及其规范化问题。

收稿日期：2013-09-02

作者简介：*李　生 (1943-)，男，黑龙江兰西人，教授，博士生导师，中国中文信息学会理事长，主要研究方向为自然语言处理、机器翻译、信息检索、社会计算等，Email：lisheng@hit. edu. cn。

针对这些问题，相关专家已经做了大量卓有成效的工作，这些工作包括：

1) 理论、方法与技术研究，结合汉语特点，引进国外技术，进行自主研发；

2) 实验和应用系统的研制开发，包括汉字处理、中文文本处理、中文语音处理、少数民族语言处理等；

3) 资源建设，包括词典和语料库等；

4) 评测，国内：863；国际：SIGHAN（分词）、NIST（机器翻译）、TREC（信息检索）等。

自然语言处理当前研究的特点：

1) 使用语料库处理大规模真实文本；

2) 使用机器学习的方法自动获取语言知识；

3) 使用统计数学（概率统计）的方法来分析语言数据；

4) 以语言知识为核心的多种方法融合。

1 自然语言处理的研究

人类自然语言的表述通常有语音和文字两种形式，本文从文字表述的角度来论述自然语言处理。自然语言处理研究包括基础研究、共性技术和应用研究，应用研究当前主要有机器翻译、信息检索和社会计算等几个方面。

1.1 基础研究

自然语言有意义的基本单元是词，按照一定的句法规则将词组织在一起就成为句子，再由句子组成段落，由段落构成篇章。自然语言处理的基础研究主要包括词法分析、句法分析、语义分析、语用语境与篇章分析等的研究。

1.1.1 词法分析

词是组成句子的基本单元。词法分析是要先将构成句子的字符串变成词串，然后再给句子中的每个词加上句法范畴标记（有时还需加上语义范畴标记）。

汉语是以汉字为单位的缺少严格意义形态变化的表意文字。

汉字处理技术包括汉字编码、汉字输入和汉字输出，汉语的句子在进行处理之前首先要切分成一个个单词，机器自动分词的方法有最大匹配法和最大概率法两种。

对于像英语这样的屈折性语言，由于词形的变化，可能造成一个词根对应着多个不同的字符串形式，在词法分析时需对其前缀、后缀及词尾等进行适当的处理，还原词形。

词性标注是词法分析的主要任务，词性是词汇基本的语法属性，也称之为词类。词性标注就是在给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程。标注的重点是解决兼类词和确定未登录词的词性问题。通常有基于规则和基于统计的两种方法。

词义标注也是词法分析的主要任务之一，重点是解决如何确定多义词在具体语境中的义项问题。对于多义词来说，一个词可以表达一个以上的意义，但它在具体的语境当中，意思往往是确定的。

标注过程中，通常是先确定语境，再明确词义。方法可仿照词性标注，有基于规则和基于统计的做法。

1.1.2 句法分析

句法分析的任务是确定句子的句法结构，识别组成句子的各个成分，明确它们之间的相互关系。判断输入的单词序列（一般为句子）是否合乎给定的语法，分析出合乎语法句子的句法结构。将能够完成这种分析任务的程序模块称之为句法分析器。

词的构成和变化规律称为词法，句子和短语的构成规则成为句法，语法研究的是语言结构的规律。狭义的语法等同于句法，广义的语法应为词法、句法、语义与语用的总称。

句法结构的形式化描述方法通常有两种：句法结构树、依存关系图。前者描述了句子的组成成分及各个成分之间的结构关系，后者则描述了句子中词与词的依存关系。

任何句子都由关键成分（主、谓、宾）和修饰成分（定、状、补）构成，关键成分为主，修饰成分为辅。通常主语和宾语为名词或代词，谓语则为动词。谓语动词在句子中处于中心地位。

句法分析通常有完全句法分析和浅层句法分析两种，前面所述的方法是指完全句法分析，要通过一系列的句法分析过程，最终得到一个句子的完整的句法树。方法也有基于规则和基于统计之分，目前以基于统计的方法为主流，概率上下文无关文法（Probabilistic Context-Free Grammar，PCFG）用得较多。完全句法分析的难点有两条：一是词性歧义；二是搜索空间太大，通常为句子中词的个数 n 的指数级。

浅层句法分析也叫部分句法分析或语块分析，它只是要求识别出句子中某些结构相对简单的成分，如非递归的名词短语、动词短语等。这些被识别出来的结构称之为语块（chunk），语块是一种介于词汇和句子之间的具有非传递特征的句子的重要成分，有时也可能就是通常的短语。这样，浅层句法分析可以分成两个子任务，一是语块的识别和分析，二是语块之间依存关系的分析。浅层语法分析的任务主要是前者。

1.1.3　语义分析

语义分析是指根据句子的句法结构和句子中每个实词的词义推导出能够反映这个句子意义（即句义）的某种形式化表示。也就是将人类能够理解的自然语言转化为计算机能够理解的形式语言。

句子的分析与处理过程，有的采用“先句法后语义”，但多数采用“句法语义一体化”的策略。目前语义分析技术还不十分成熟，近年出现一些运用统计方法获取语义信息的研究，常见的有词义消岐和浅层语义分析。

词义消岐的基本思路是对每个需要消岐的多义词先寻找出所在的上下文特征，根据这个特征来确定在特定的语境中词义的选择（在前面的词义标注中已有阐述）。

浅层语义分析又称语义角色标注，语义角色标注是将句子中的句法成分标注成为谓语动词的语义角色，并将每个语义角色赋予一定的语义含义。语义角色通常是与句子中的某一句法成分相对应的，如施事——主语，受事——宾语，时间——时间状语，地点——地点状语等。

如对于句子“校学位委员会昨天在邵馆讨论了博士学位授予问题。”这里动词“讨论”为谓语动词，“校学位委员会”是施事，“博士学位授予问题”是受事，“昨天”是发生的时间，“邵馆”是发生的地点。形式表示为：讨论（校学位委员会，博士学位授予问题，昨天，邵馆）。这里用短语（也就是浅层句法分析中所说的语块）作为浅层语义分析的基本单元。这样，浅层语义分析就可以建立在短语结构句法分析或依存句法分析的基础之上。

语义角色标注通常被看成是分类问题，标注的步骤一般为：剪枝——去掉不重要的句法成分；识别——找出可承担语义角色的句法成分；分类——进行具体的语义角色分类与标注；后处理。支持向量机 SVM，最大熵 ME，决策树 c4.5 的改进随机森林算法和 SNow 算法等都曾成功地运用到语义角色标注上。

基础研究中还应包括语用语境与篇章分析。语用是指人对语言的具体运用，研究和分析语言使用者的真正用意，它与语境、语言使用者的知识状态、言语行为、想法和意图有关联，是对自然语言的深层理解。语境分析主要涉及的是情景语境和文化语境。篇章分析是将研究扩展到句子界限之外，对段落和整篇文章进行理解和分析。

除此之外还需对词义消岐（确定在给定上下文语境中多义词的词义）、指代消解（确定指代词的先行语的过程）、命名实体识别（人名、地名、组织机构名、数量表达式、时间短语、货币短语和百分比等的识别）等方面进行研究。

1.2　共性技术

共性技术是指在自然语言处理当中经常用到的技术，这些技术有时也可以独立应用。

1.2.1　文本分类与聚类

采用机器学习方法，依据自然语言文本属性的相似程度来对文本进行存储与管理。文本分类是一个有指导的学习过程。它根据一个已经被标注的训练文本样本集合，找到文本属性和文本类别之间的关系模型，然后利用这种学习得到的关系模型对新的文本进行类别判断。

文本聚类是一个无指导的学习过程。它是根据文本数据的不同特征，将其划分为不同数据类的过程，其目的是使同一类别的文本间的距离尽可能

小，而不同类别的文本间的距离尽可能的大。

1.2.2 信息抽取与文本挖掘

信息抽取是指从自然语言文本中抽取特定信息 (包括实体、关系、事件等)，经过结构化处理变成表格形式之后再对信息进行存储和管理的过程。

文本挖掘是指从大量文本集合中获取用户感兴趣或者有用的模式的过程，挖掘是从杂乱无章的数据中寻找知识，在目前所处的大数据时代，数据挖掘和机器学习尤为重要。

1.2.3 自动文摘

文摘是依据用户需求从源文本中提取最重要的信息内容，生成一个精简版本的过程。文摘应具有压缩性、内容完整性和可读性。文摘可分为单文档文摘和多文档文摘。运用计算机自动生成的文摘称之为自动文摘，自动文摘的生成有浅层方法和深层方法。依据生成方法的不同又可分成机械式文摘和理解式文摘，目前自动文摘的生成多半是采用机械式方法，理解式文摘是建立在对自然语言的理解的基础之上的，难度较大，实现起来还有一定困难。

1.2.4 复述与文本生成

复述研究的是短语或句子的同义现象，任务有两条：一是识别两个短语或句子是否互为复述——抽取，二是将给定的短语或句子复述成另外一个短语或句子——生成。复述保留了“概念上的近似等价”，而结构却不一定相似。

文本生成是研究计算机如何根据信息在机器内部的表达形式生成一段高质量的自然语言文本。

1.2.5 话题检测与跟踪

在海量数据流中自动发现话题，并将与话题相关的内容联系在一起。时间是话题的一个重要特征，从时间概念出发，话题又可以分成“突发性话题”和“持久性话题”。话题具有“语义”和“时间”两个主要特征。除了事件内容之外，话题还通常包涵人物、时间、地点等命名实体。

1.2.6 情感分析

识别出文本中所包含的主观性句子，并对其情感趋势进行分析与判断。

例：我前几天买了一部手机，它不仅外观漂亮，而且性能很好。

这里的第一个句子为客观句，二、三两句是主观句，主体是“我”，主题（评价对象）分别是“外观”和“性能”，它们都是手机的属性，而情感词（评价词）分别为“漂亮”和“很好”。

处理过程大致如下：

识别出主观句—找出主题词—识别出情感词—判断出情感词的极性—句子倾向性分析—确定主体。

1.2.7 语料库与词汇知识库

自然语言处理的资源建设主要是指语料库与知识库的建设。语料库是存放语言材料的仓库，自然语言处理领域的语料库则是按照一定原则组织在一起的大规模真实自然语言数据的集合。要求：库存要有一定的规模；应为实际使用中的真实语言材料（书面语或口语）；需经分析、加工、处理（标注）。语料库主要用于研究自然语言规律，特别是统计语言学模型的训练及相关系统的评价与评测。

语料库根据它所包含的语言种类数目分为单语语料库和多语语料库，双语语料在使用时往往还需要进行双语对齐。

国际上最有代表性的是词汇知识库由美国普林斯顿大学认知科学实验室 George A. Miller 领导的研究组开发的英语机读词汇知识库 WordNet，它是以同义词集合做为基本的建构单位，给出了同义词集合的定义和例句。国内代表性的成果是由董振东父子创建的汉英双语语言知识库知网 HowNet，它是以概念为描述对象，以揭示概念与概念之间以及概念与所具有的属性之间的关系为基本内容的常识知识库。

1.3 应用研究

1.3.1 机器翻译

机器翻译是运用计算机来实现不同语言之间的自动翻译。通常，被翻译的语言称之为源语言，翻译结果的语言称之为目标语言。机器翻译就是从源语言到目标语言的转换过程。从形式上看，机器翻译是一个符号序列的变换过程。机器翻译方法总

体上可以分成基于规则的和基于语料库的两大类。

1) 基于规则的机器翻译方法。使用的主要资源是词典与知识库（存放规则与常识性知识）。又可分成基于转换的和基于中间语言的两种方法。

基于转换的方法通常由分析、转换、生成 3 个步骤构成。这里的分析是指对源语言句子的分析，包括词法分析、句法分析、语义分析、语境分析等等，重点在句子的结构分析，经过分析之后生成源语言的句法结构树（往往附有一定的语义信息）；转换阶段要依据翻译规则实现将源语言的句法结构树转换成等价的目标语言的句法结构树；再运用词典和常识性知识等完成目标语言的生成。在实际翻译中往往是一个由词到短语再到句子的分层次转换的过程。

基于中间语言的方法要先将源语言句子转换成一种与具体语种无关的通用语言或中间语言，然后再将这种语言的句子转换成目标语言的句子。整个翻译过程包含两个独立转换的过程。该方法适用于一对多的翻译，基于枢轴语翻译属于这种方法。

2) 基于语料库的机器翻译方法。使用的主要资源是经过标注的语料库，语料库是按照一定原则组织在一起的大规模真实自然语言数据的集合。又可分成基于实例的和基于统计的两种方法。

基于实例的方法需要对已有的语料进行词法、句法甚至语义等分析，建立存放翻译实例的实例库。系统在执行翻译的过程中，将翻译句子与实例库中的翻译实例进行相似性分析，其中最相似的句子的译文便为翻译句子的译文。

基于统计的方法是以大规模双语对齐语料库为基础，对源语言和目标语言词汇的对应关系进行统计，通过词汇同现的可能性来计算两种语言之间词汇映射的概率，据此产生目标语言的译文。它是用机器学习的方法来解决机器翻译中的问题。

统计机器翻译有 3 个基本问题：一是建模，通过数学模型来描述翻译过程；二是训练（学习），利用双语语料库估计模型参数；三是解码（搜索），利用已获得参数的模型，对给定输入的源语言句子，找出最优（概率最大）的目标语言句子候选。统计机器翻译可以表示为

$$E=\arg\max P(E)\times P(E|F),$$

其中，$P(E|F)$为翻译模型，计算源语言翻译成目标语言的概率，反映的是两个句子互为翻译的可能性，达到较好的忠实度；$P(E)$为语言模型，反映的是句子在目标语言中出现的可能性，也就是看其在语法语义等方面（主要是语法）的合理程度，即看是否有一个较好的流利度。$P(E)$只与目标语言有关，重点解决调序和搭配问题，可采用n元文法和链语法等语法模型。

目前机器翻译经常使用的方法有基于规则的（实际上是指基于转换的），基于实例的和基于统计的 3 种。基于规则的方法通过计算机程序最好地反映了人们对于语言翻译的认知和理解，基于实例的方法有效地发挥了计算机的存储能力，而基于统计的方法充分发挥了计算机的数学建模能力。

市场上的机器翻译系统多半是基于规则的和基于实例的，但由于基于规则的机器翻译系统人工编写规则的工作量太大，知识库的规模和一致性都难以把握。基于实例的机器翻译系统的不足在于翻译实例的泛化、覆盖率以及实例的匹配等问题。基于统计的机器翻译大规模细粒度知识的自动获取能力较强，可以弥补前面两者的一些不足。基于统计的机器翻译方法已成为当前的主流研究方向。

影响机器翻译系统质量的主要障碍是歧义问题的处理和常识性知识的使用。

1.3.2　信息检索

信息检索是指从有关文档集合中查找用户所需信息的过程。广义的信息检索是指先将信息按一定的方式组织和存储起来，然后再根据用户的需求从已经存储的文档集合当中找出相关的信息。其中包括“存”与“取”两个方面，“存”即信息存储，是对信息进行收集、标引、描述、组织，进行有序的存放。“取”即信息查找，是按照某种查询机制从有序存放的信息集合（数据库）中找出用户所需信息或获取其线索的过程。

信息检索的基本原理是将用户的检索提问词（关键词）与数据库文献纪录中的标引词进行对比，二者匹配一致时，即为命中，检索成功。这里“存”和“取”的联系一致是通过检索标识来实现的，检索标识是为沟通文献标引和检索提问而编制的人工语言。检索结果按与提问词的关联度输出，供用

户选择。用户通常是采用“关键词查询+选择性浏览”的与机器交互方式获取信息。

信息检索最早是在 20 世纪 50 年代提出的，90 年代互联网出现以后，其导航工具——搜索引擎可以看成是一种特殊的信息检索系统，如果说二者有区别的话，那就是语料库集合和用户群体有所不同，搜索引擎面临的语料库是规模浩大、内容繁杂、动态变化的互联网，用户群体不再是具有一定知识水平的科技工作者，而是兴趣爱好、知识背景、年龄结构差异很大的网民群体。目前多数人习惯于二者通用。

以谷歌为代表的“关键词查询+选择性浏览”的交互方式的特点是：用户用简单的关键词作为查询提交给搜索引擎，搜索引擎并非直接把检索目标页面反馈给用户，而是提供给用户一个可能的检索目标页面列表，用户浏览该列表并从中选择出能够满足其信息需求的页面加以浏览。

这种交互方式对于用户来说查询输入简单了，但机器却难以通过简单的关键词准确的理解用户的真正查询意图，因此只能将有可能满足用户需求的结果集合以列表的形式提供给用户。

目前互联网是人们获取信息的主要来源，网络上存放着取之不尽、用之不竭的信息，网络信息有着海量、分布、无序、动态、多样、异构、冗余、质杂、需求各异等特点。人们现在已经不再满足于目前的搜索引擎带来的查询结果，对下一代搜索引擎的要求是个性化（精确化）、智能化、商务化、移动化、社区化、垂直化、多媒体化、实时化等诸多方面，从事这方面研究的专家们正在努力满足人们日益增长的需求。

要机器自动地从互联网上找出问题的答案是许多用户提出的新的需求，问答系统可以满足这一需求。这里的用户查询（提问）是自然语言，而返回的结果是直接答案（不再是网页）。

许多专业用户还要求搜索引擎能够推荐所需信息，信息推荐（过滤）系统便可以完成这一任务。信息检索是针对动态变化的信息需求从固定的信息集合中获取相关知识，信息过滤则是针对固定的信息需求从动态变化的信息流中获取相关知识。二者都是依靠信息的相关性进行判断。

衡量信息过滤效果的依据在于系统要尽可能多地获取相关信息，而同时也要尽可能多地屏蔽掉不相关信息。这里的关键技术在于去噪声能力要强。噪声既来源于不相关文本，也来源于相关文本中的不相关信息。

1.3.3 社会计算

社会计算也称计算社会学，是指在互联网的环境下，以现代信息技术为手段，以社会科学理论为指导，帮助人们分析社会关系，挖掘社会知识，协助社会沟通，研究社会规律，破解社会难题的学科。社会计算是社会行为与计算系统交互融合，是计算机科学、社会科学、管理科学等多学科交叉所形成的研究领域。它用社会的方法计算社会，即是基于社会的计算，也是面向社会的计算。

社会媒体是社会计算的主要工具和手段，它是一种在线交互媒体，有着广泛的用户参与性，允许用户在线交流、协作、发布、分享、传递信息，组成虚拟的网络社区等等。近年来，社会媒体呈现多样化的发展趋势，从早期的论坛、博客、维基到风头正劲的社交网站、微博和微信等，正在成为网络技术发展的热点和趋势。社会媒体文本属性特点是其具有草根性，字数少、噪声大、书写随意、实时性强；社会属性特点是其具有社交性，在线、交互。它赋予了每个用户创造并传播内容的能力，实施个性化发布，社会化传播，将用户群体组织成社会化网络，目前典型的社会媒体是 Twitter 和 Facebook，在我国则是微博和微信，用户已经超过了 3 亿。微博即微博客 (Micro blog) 的简称，是一个基于用户关系的信息分享、传播以及获取的平台。

社会媒体是允许用户广泛参与的新型在线媒体，通过社会媒体用户之间可以彼此之间在线交流，形成虚拟的网络社区，构成了社会网络。社会网络是一种关系网络，通过个人与群体及其相互之间的关系和交互，发现它们的组织特点、行为方式等特征，进而研究人群的社会结构，以利于他们之间的进一步共享、交流与协作。

2 自然语言处理的发展展望

中国中文信息学会创建于 1981 年，30 多年来取得了诸如汉字激光照排、联想汉卡、汉王手写输

入、亚伟速录、科大讯飞语音合成、TRS 中文检索、北大语法信息词典、知网以及自然语言处理的基础研究与共性技术、机器翻译、信息检索等应用研究等多项高水平的研究成果。随着现代信息技术的高速发展，随着我国综合国力的不断提高和对外交往的不断加大，党和国家对信息技术，对自然语言处理特别是中文信息处理更加重视，相关部门也更加支持。党的十七大、十八大及国家中长期科学和技术发展纲要等都提出了明确要求，自然语言处理相关研究一定会有更大的发展。本人对未来的信息技术、自然语言处理的发展及从事这方面研究的工作进行了一些思考。

2.1　物联网与人联网（社会网络）

最近几年信息领域的理论和技术发展迅速，从并行计算、分布式计算、网格计算到云计算、物联网，从 Web 网、社会媒体、社会网络到社会计算，从数据库、数据仓库到大数据，计算技术与互联网特别是移动互联网的结合大大推进了信息技术的发展，信息技术正在改变未来。

物联网是实现物物相连的互联网络。是指将各种识别及传感设备，如 RFID、GPS、传感器、红外感应器和激光扫描仪等嵌入到物体当中，按照约定的协议，再将这些物体用无线或有线通信网络连接起来，所形成的人与人、物与物、人与物之间可以广泛进行信息沟通的新型网络。

社会网络是通过社会媒体实现用户（人与人）之间的在线交流，形成虚拟的网络社区。社会网络是一种关系网络，通过个人与群体及其相互之间的关系和交互，发现它们的组织特点、行为方式等特征，进而研究人群的社会结构，以利于他们之间的进一步共享、交流与协作。

未来以信息技术为核心的信息空间将会把人类社会和物理世界更加紧密地联系在一起，实现人类社会、信息空间、物理世界三者的全面连通与融合。信息空间与人类社会的关系将由以技术为中心转化为以服务为中心。为了实现服务的普适化（无所不在，随时随地），要将感知设备（如 RFID，传感器等）和计算设备嵌入到物理世界的实体（人和物）中去，再由泛在网将它们全面的连通。

在网络环境下，作为社会主体的人通过具有共同的兴趣、爱好、价值及行为等特征相互联系在一起构成网络的虚拟社区，形成虚拟社会。虚拟社会是现实社会的映射，它与现实社会相互关联、相互影响，现实社会的矛盾与问题能够很快地反映到虚拟社会中，并能迅速的传播与扩展，当然也可以利用虚拟社会去化解和沟通。未来社会的进步和发展离不开物联网，也离不开人联网。

2.2　拓宽思维、扩大视野，抓住发展时机

自然语言处理的研发要充分运用语言学知识，依靠云计算和大数据，获取资源，运用工具，掌握新的对策与方法。

信息技术要把握住感知、连通与计算 3 个要素，与商业、社会科学、管理科学、心理学、认知科学及生命科学等，当然也包括数学和物理学的交叉与融合。

把握并运用好计算思维和网络思维，求解复杂问题由只靠专家到专人，更要注重网络上的用户群体。由事后处理转化成更多的事先预测。科技工作者要具备学术的敏感性，紧紧把握住学术的动态变化、发展趋势和前沿问题。

牛津大学的维克托提出大数据环境下思维大变革的 3 个转变：不要抽样要全体；不要精确要混杂；不要因果要相关。确切一点说，应该是在求解复杂问题时更要注重后者。

大数据的绝大部分是非结构化数据，大数据更多来源于人的言行，大数据与自然语言处理密切相关，大数据进一步缓解了统计机器学习方法中的数据稀疏问题。

要紧紧把握住泛在 (无所不在，无所不能)、服务（以服务为中心，创新要更多地体现在商业运作模式上）和人本（基于人，为了人，用社会计算社会，基于与面向）3 个计算；做好感知连通与计算、软件与硬件、理论技术与服务 3 个深度结合；实现计算功能由数值计算到信息处理的转变，计算方法由单机到多机分布计算网格计算直至云计算的转变，开发应用由以技术为中心到以服务为中心的转变（商业运作模式），知识获取由依赖专家到依赖用户（专家—专人—自然标注—众包）的转变，研究方法由以模型为重点到以（大）数据为重点的转变、信息处理由表层特征（文本—语法，图像—形

状、颜色、纹理等）向深层语义分析转变，由用户采用关键词搜索到个性化推荐，直接回答用户问题等转变。

2.3 从需求出发，把握好研究方向

处理好计算服务、计算技术与计算科学三者之间的关系。从需求出发，为需求服务，从中提升科学问题。

科学研究的对象是领域中的共性问题。往往要先对现实领域中的个性问题进行“白盒研究”，知识积累多了就有可能抽象出领域通用的共性问题，再研究其“黑盒模型”及普适规律。

既要考虑到发展的需求，也要考虑到应用的需求。研究的问题要有价值、有意义，目的性和必要性要清楚。需求—应用问题—科学问题—实施方案—预期效果 。学术（研究）要跟着工程走。

经济与社会的发展进一步扩大了对自然语言处理研究的需求：国内外交往的信息及语言支持的需求；互联网上海量信息的处理的需求；国家文化产业振兴的支柱的需求；经济与社会发展的保障的需求；缩短与国际前沿差距的需求等等。

2.4 充分利用资源，建设好学术基地

要充分利用互联网、大数据及海量信息资源；要充分利用（超大）计算机的速度和容量资源；要充分利用云服务（计算与存储）资源；要充分利用现有的资源、工具、技术与方法；要充分重视基础研究与共性技术开发；要充分调动人的智慧与才能；要充分进行合作与交流；要充分考虑市场需求（以服务为中心）。还要建设好基础理论与共性技术研究基地；推广应用与工程技术开发基地和交流合作与人才培养学术基地。

抓住机遇，乘势而上，信息技术和自然语言处理一定会大有作为，一定会为经济与社会发展、为民族振兴做出更大的贡献。

参考文献

[1] 2006-2020 年国家信息化发展战略 [Z]. 中共中央办公厅、国务院办公厅, 2006.

[2] 宗成庆. 统计机器翻译 [M]. 2 版. 北京: 清华大学出版社, 2013.

[3] 刘奕群, 马少平, 洪涛, 等. 搜索引擎技术基础 [M]. 北京: 清华大学出版社, 2010.

[4] 刘挺, 秦兵, 张宇, 等. 信息检索系统导论 [M]. 北京: 机械工业出版社, 2008.

[5] 王晓龙, 关毅. 计算机自然语言处理 [M]. 北京: 清华大学出版社, 2005.

[6] 冯志伟. 计算语言学基础 [M]. 北京: 商务印书馆, 2001.

[7] 维克托·迈尔-舍恩伯格, 肯尼迪·库克耶. 大数据时代：生活、工作与思维的大变革 [M]. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013.

[8] Lazer D, Pentland A, Adamic L, et al.. Computational Social Science [J]. Science, 2009,323 (5915): 721-723.

[9] Zhang Hui, Zhang Min, Li Haizhou, et al.. Fast translation rule matching for syntax-based statistical machine translation [C] // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2, Singapore, 2009: 1037-1045.

[10] Hirschman L, Gaizauskas R. Natural language question answering: the view from here [J]. Natural Language Engineering, 2001,7 (4): 275-300.

Research and development of natural language processing

LI Sheng

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Natural Language Processing (NLP) is one of the key issues in the information technology. NLP for Chinese covers a broad topics of processing the character, the word, the sentence, the paragraph and the discourse. In this paper state-of-the-art NLP researches is summarized, with a special focus on the unique challenges in Chinese NLP. The achievements and existing problems in NLP is reviewed, and its possible future work is outlined as well from a personal perspective.

Key words: natural language processing; Chinese information; common technology; prospect