自然语言处理

1.问题的提出

- 自然语言是指人类日常使用的语言，如汉语、英语等。
- 语言是思维的载体，是人类交流思想、表达情感最自然、最直接、最方便的工具。
- 人类历史上以语言文字形式记载和流传的知识占知识总量的80%以上。

- 无处不在的网络、无处不在的通讯、堆积如山的文档，构成了当今社会信息爆炸的基本特征。当现代化的信息传播手段给人们的生活和工作带来极大便利的同时，也使人们面临许多难以克服的困难和障碍。有关专家指出，语言障碍是21世纪社会全球化所面临的主要困难之一。

2.基本概念

定义：自然语言处理

“自然语言处理可以定义为研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力和语言应用的**模型**，建立计算**框架**来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实用**系统**，并探讨这些实用系统的评测技术。”

美国计算机科学家Bill Manaris　　《计算机进展》

定义：自然语言处理

自然语言处理就是利用计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。

- 冯志伟，《自然语言的计算机处理》

定义：自然语言理解(Natural Language Understanding, NLU)

自然语言理解是人工智能早期的研究领域之一。从微观上讲，语言理解是指从自然语言到机器(计算机系统)内部之间的一种映射。从宏观上讲，语言理解是指机器能够执行人类所期望的某些语言功能。这些功能包括回答有关提问、提取材料摘要、不同词语叙述、不同语言翻译。

- 蔡自兴、徐光佑，《人工智能及其应用》
清华大学出版社，2004

关于“理解”的标准

图灵(Turing)测试：

定义：计算语言学(Computational Linguistics)

计算语言学是利用电子数字计算机进行的语言分析。虽然许多其他类型的语言分析也可以运用计算机，计算分析最常用于处理基本的语言数据 - 例如，建立语音、词、词元素的搭配以及统计它们的频率。

- 《大不列颠百科全书》

计算语言学是语言学的一个研究分支，用计算技术和概念来阐述语言学和语音学问题。已开发的领域包括自然语言处理，言语合成，言语识别，自动翻译，编制语词索引，语法的检测，以及许多需要统计分析和领域（如文本考释）。

－戴维•克里斯特尔，《现代语言学词典》

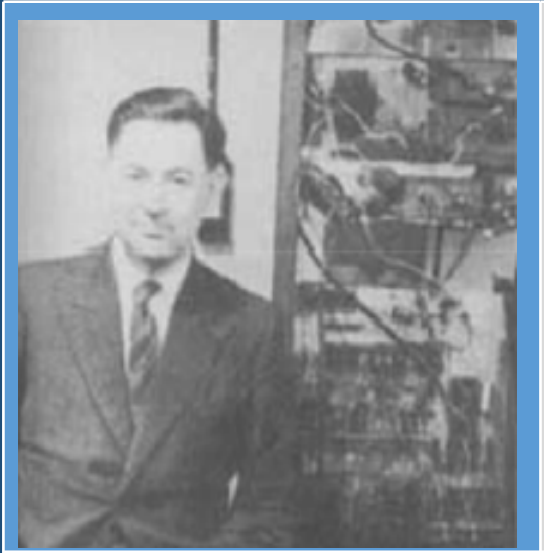
近几年来，自然语言处理技术迅速发展成为一门相对独立的学科，倍受关注，而且该技术不断与语音技术相互渗透和结合形成新的研究分支，因此，很多人在谈到“计算语言学”、“自然语言处理” 或“自然语言理解”这些术语时，往往默认为同一个概念。

汉语已经不再只是我们自己使用和关注的语言，外国人可能喜欢她或者讨厌她，但不敢藐视她！

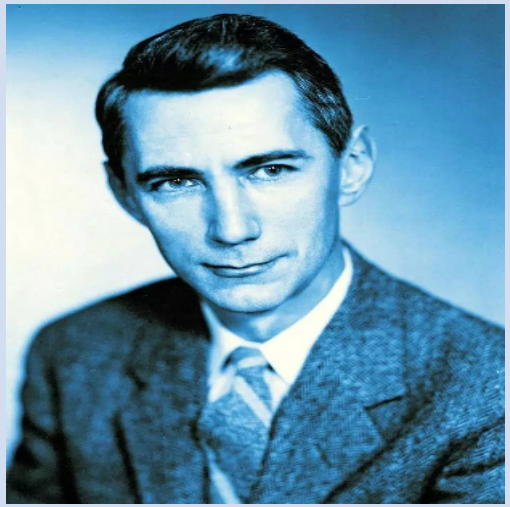
针对汉语的处理技术早已成为学术界和企业界共同关注的问题，汉英两大强势语言的自动翻译问题则是人类语言技术中最具挑战性的研究课题。

3.NLP的产生与发展

- 源自机器翻译 (Machine Translation, MT)
- 英国 Andrew Donald Booth 和美国W. Weaver 提出机器翻译的概念。
- 随着机器翻译研究的进展，各种自然语言处 理技术应运而生，并逐渐发展壮大，形成了这一语言学与计算机技术相结合的新兴学科。



- A. D. Booth，数学家、物理学家，曾研究利用X射线确定晶体结构，二战中参与计算机研制，在程序化计算机研究中成绩卓著；



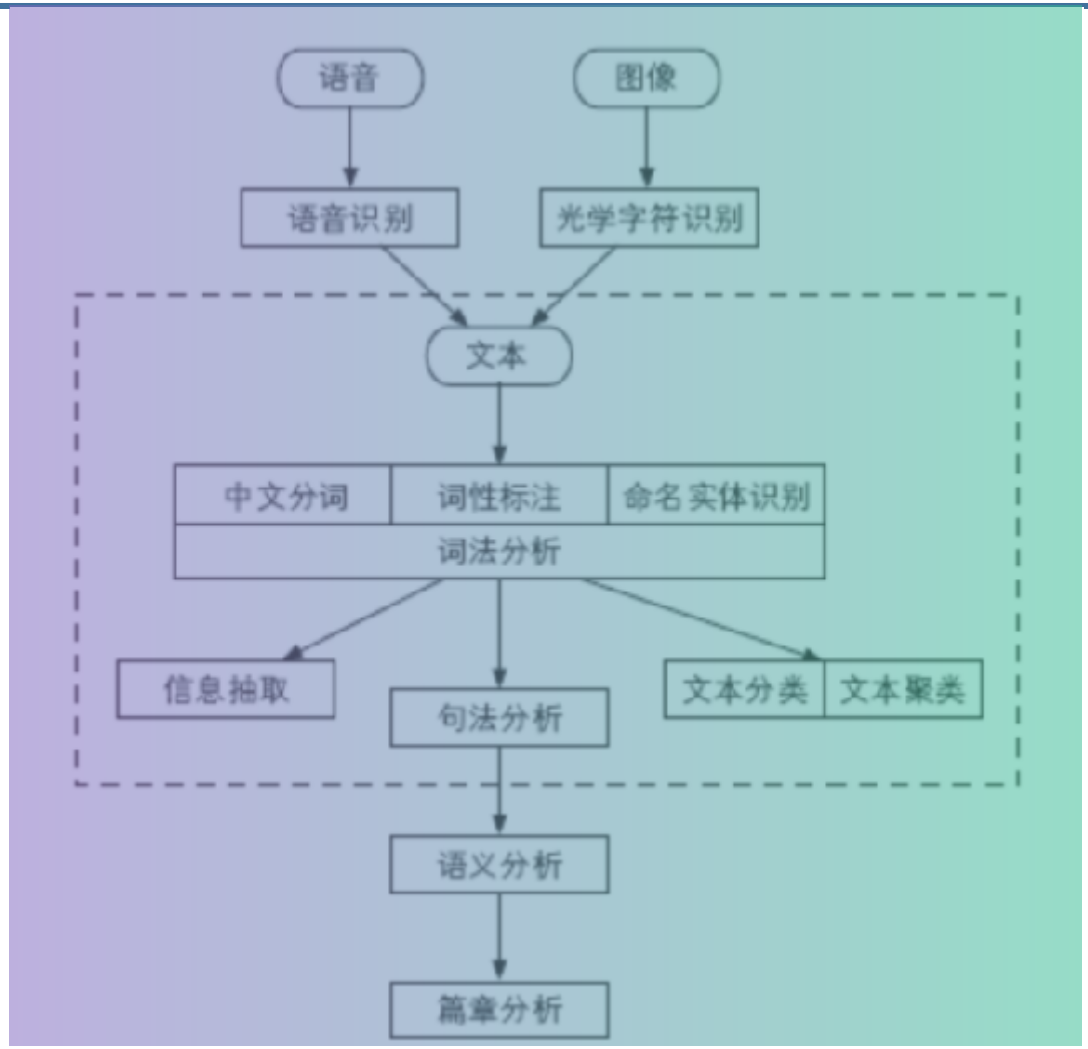
- 香农，信息论先驱
- 1920至1932年曾在Wisconsin大学 教授数学；
- 1932至1955年担任Rockefeller Institute自然科学部主任。

曲折的发展历程：

- 1960S 中期之前：萌芽期
- 1960S 中期到1970S 中后期：步履维艰

  — 1966年美国科学院发表 ALPAC报告
- 1970S 中后期到1980S 后期：复苏
- 1980S 后期至今：蓬勃发展

4.NLP的层次


语音
图像
语音识别
光学字符识别
文本
中文分词
词性标注
命名实体识别
词法分析
信息抽取
句法分析
文本分类
文本聚类
语义分析
篇章分析

5.NLP的基本任务

5.1 语音识别

1.定义

- 语音识别也称为自动语音识别（Automatic Speech Recognition，ASR），它是以语音为研究对象，通过语音信号处理和模式识别让机器理解人类语言。

2.语音识别的应用场景

文字输入	录音整理	聊天机器人
语音转换	语音检索	智能声控
会场速记	字幕转换	人机交互

5.2 光学字符识别

1. 定义

- **光学字符识别（Optical Character Recognition，OCR）是利用光学和计算机技术将纸质文档中的文本转换成图像文件，然后利用图像处理技术和字符识别算法辨别图像文件上的字符，并将所获信息存在计算机文件中的一种技术。**

2.语音识别的应用场景

证件识别

包括身份证识别，护照识别，名片识别和银行卡识别。

1

发票识别

包括手写支票识别，内容转换成数字文本和验证签名等，适合于银行、税务等行业大量票据表格的自动扫描识别及长期存储。

2

车牌识别

车牌识别技术在智能交通，小区停车场等都有很好的应用，车牌识别就是对车牌进行OCR识别，再进行比对的过程。

3

5.3 中文分词

1. 定义

- **中文分词就是是在中文句子中的词与词之间加上边界标记。为了让计算机更容易理解文本，通常中文信息处理的第一步是中文分词。**

2.中文分词存在的困难

交集型歧义

指的是字符串“ACB”中，“AC”和“CB”都是一个汉语词汇，会存在多种切分交叉在一起。例如：

在“研究所有东西”中，“研究所/有/东西”或者“研究/所有/东西”都是合理的切分方式。

1

组合歧义

指的是字符串“AB”是一个词汇，“A”和“B”同时也是词汇，存在不同语义下切分不同。

“这个门把手坏了”　　“请把手拿开”

2

5.4 词性标注

1. 定义

- **词性标注（Part-Of-Speech tagging，POS tagging）是将句子中的每一个词的词性按其含义和上下文内容进行标记的文本处理技术，也就是确定每个词是名词、动词、形容词等词性的过程。**

5.4 词性标注

2. 宾州中文树库标注集中的汉语词性标注示例：

原句：这当然不是历史的巧合，而是历史的积累和转接。

词性标注后：这/PN 当然/AD 不/AD 是/VC 历史/NN 的/DEG 巧合/NN ，/PU 而/CC 是/VC 历史/NN 的/DEG 积累/NN 和/CC 转接/NN 。/PU

5.5 命名实体识别

1. 定义

- **命名实体识别（Named Entity Recognition，NER），指的是文本中具有特定意义实体或实体指代项。命名实体识别通常指的是在文本中标识命名实体并划分到相应的实体类型。显而易见，命名实体识别包括实体的边界识别和确定实体的类型两个子任务。**

广义的命名实体分为三大类（实体类，时间类和数字类），若干小类（人名、地名、组织机构名、时间、日期等），在研究中，命名实体识别通常指的是狭义的命名实体，即人名、地名、组织机构名。

2.命名实体识别的困难

1.中文命名实体需要预先分词。分词的错误在命名实体识别过程中无法得到纠正，错误蔓延会影响到命名实体识别。

2.相比英文，中文中没有大小写及词形变换特征。

3.命名实体表述多样，没有可以共同遵循的严格命名规范。

4.未登录词问题。

5. 中文命名实体识别的歧义问题。

- 分类歧义。如“小米”既可以代表一种粮食作物，也可以表示为一家科技公司。
- 分词歧义指的是根据命名实体分界的不同，可以有不同的结果。比如：“李府正在吃饭”，可以被分成“李府/正在/吃饭”，也可以被分成“李府正/在吃饭”。

6.命名实体识别的领域局限性。命令实体识别在某个领域取得较好的成绩，如新闻。由于不同领域的数据往往具有领域独特特征，这些技术很难迁移到其他特定领域中，如军事、医疗、小语种语言等。

5.6 句法分析

1. 定义

- 句法分析的主要任务是识别出句子所包含的句法成分以及这些成分之间的关系。
- 句法分析将输入句子从**序列**形式变为**树状**结构，从而可以捕捉到句子内部词语之间的**远距离搭配或修饰关系**。
- 目前研究界存在两种主流的句法标注体系，句法结构分析和依存句法分析。