# 自然语言处理与生物识别

Natural Language Processing for Biometric Identification

2020 年 2 月 18 日

Springer

# 目录

# Chapter 1
# NLP 发展现状与应用领域

## 1.1 定义

自然语言处理（Natural Language Processing，简称 NLP）是用计算机来处理、理解以及运用人类语言，它属于人工智能的一个分支，是计算机科学与语言学的交叉学科，又常被称为计算语言学。由于自然语言是人类区别于其他动物的根本标志；没有语言，人类的思维也就无从谈起。所以自然语言处理体现了人工智能的最高任务与境界；也就是说，只有当计算机具备了处理自然语言的能力时，机器才算实现了真正的智能。

从研究内容来看，自然语言处理包括语法分析、语义分析、篇章理解等。从应用角度来看，自然语言处理具有广泛的应用前景。特别是在信息时代，自然语言处理的应用包罗万象，例如：机器翻译、手写体和印刷体字符识别、语音识别及语音合成、信息检索、信息抽取与过滤、文本分类与聚类、舆情分析和观点挖掘等，它涉及与语言处理相关的数据挖掘、机器学习、知识获取、知识工程、人工智能研究和与语言计算相关的语言学研究等。

自然语言处理研究的问题一般会涉及自然语言的形态学、语法学、语义学和语用学等几个层次。

形态学（morphology）：形态学是语言学的一个分支，重点研究词的内部结构，包括屈折变化和构词法两个部分。

语法学（syntax）：研究句子结构成分之间的相互关系和组成句子序列的规则。其关注的中心是：为什么一句话可以这么说，也可以那么说?

语义学（semantics）：语义学的研究对象是语言的各级单位（词素、词、词组、句子、句子群、段落、篇章）的意义，以及语义与语音、语法、语境的关系等等。其重点在探明符号与符号所指对象之间的关系，从而知道人们的语言活动。简而言之，它关注的重点是：这个语言单位到底说了什么?

语用学（pragmatics）：大概来说它必须说明的问题是多方面的，包括直指、会话隐含、预设、言语行为、话语结构等。它研究在不同上下文下的语句应用，以及上下文对语句理解所产生的影响。其关注重点在于：为什么在特定的上下文要说这句话?

自然语言处理需要解决的关键是歧义消解问题和未知语言现象的处理问题。一方面，自然语言中存在大量的歧义现象，无论是在词法层次、句法层次，还是在语义层次和语用层次，无论哪类语言单位，其歧义性始终是困扰人们实现应用目标的一个根本问题。

## 1.2 发展历史

对于自然语言处理的发展历程，可以从哲学中的理性主义和经验主义说起。基于规则的自然语言处理是哲学中的理性主义；基于统计的自然语言处理是哲学中的经验主义。理性主义方法认为，人类语言主要是由语言规则来产生和描述的，因此只要能够用适当的形式将人类语言规则表示出来，就能够理解人类语言，并实现语言之间的翻译等各种自然语言处理任务。而经验主义方法则认为，从语言数据中获取语言统计知识，有效建立语言的统计模型。因此只要能够有足够多的用于统计的语言数据，就能够理解人类语言。

早期的自然语言处理具有鲜明的经验主义色彩。如 1913 年马尔科夫提出马尔科夫随机过程与马尔科夫模型 [1] 的基础就是手工计算长诗中元音与辅音出现的频度；1948 年香农把离散马尔科夫的概率模型 [2] 应用于语言的自动机，同时采用手工方法统计英语字母的频率。

然而这种经验主义到了乔姆斯基时出现了转变。1956 年乔姆斯基 [3] 借鉴香农的工作，把有限状态机用作刻画语法的工具，建立了自然语言的有限状态模型，具体来说就是用代数和集合论将语言转化为符号序列，

找到了一种一种统一的数学理论来刻画自然语言，从此诞生了一个叫做“形式语言理论”的新领域。

在 20 世纪 50 年代末到 60 年代中期，经验主义东山再起了。多数学者普遍认为只有详尽的历史语料才能带来可靠的结论。于是一些比较著名的理论与算法就诞生了，如贝叶斯方法、隐马尔可夫 [4]、最大熵 [5]、维特比算法 [6]、支持向量机 [] 等之类。世界上第一个联机语料库 Brown Corpus[7] 也是在那个时候的 Brown University 诞生的。但是总的来说，这个时代依然是基于规则的理性主义的天下，经验主义虽然取得了不俗的成就，却依然没有受到太大的重视。

随着 90 年代以来，基于统计的自然语言处理开始大放异彩了。首先是在机器翻译领域取得了突破，因为引入了许多基于语料库的方法。1990 年在芬兰赫尔辛基举办的第 13 届国际计算语言学会议之后，大家的重心开始转向大规模真实文本了，传统的仅仅基于规则的自然语言处理显然力不从心了。学者们认为，大规模语料至少是对基于规则方法有效的补充。到了 1994 至 1999 年，经验主义就开始空前繁荣了。如句法剖析、词类标注、指代消解等算法几乎把概率与数据作为标准方法，成为了自然语言处理的主流。下面列举一些比较重要的里程碑。

1998 年和接下来的几年里，FrameNet[8] 项目指导了语义角色标注的任务。这是一种浅层语义解析的形式，后续促进了，如组块分析、命名实体识别、依存分析等核心 NLP 任务的研究。

2001 年，条件随机场 [9]（Conditional Random Fields，CRF）成为了最具影响力的序列标注方法类别之一，获得了 ICML 2011 的最佳论文奖。CRF 是当前最先进的序列标注问题模型的核心部分，这些模型具有标签间的相互依赖性，广泛用于分词、词性标注、专名识别等序列标注任务中。

2002 年，双语互译质量评估辅助工具 BLEU[10]，给出了双语互译质量度量标准，这使得机器翻译系统得以扩展，其现在仍然是机器翻译评估的标准度量标准。同年，结构化感知机 [11] 问世，为结构化感知工作奠定了基础。同时，情感分析 [12] 也成了最受欢迎和广泛研究的 NLP 任务之一。

在 2003 年，引入了潜在狄利克雷分布 [13]，至今仍是主题建模的标准方法。在 2004 年，有学者提出了比 SVM 更适合于捕获结构化数据中的相关性的新最大边缘模型 [14]。

2006 年，发布了 OntoNotes[15] 这个大型多语言语料库，它已被用于训练和评估各种任务，如依赖解析和引用解析。在 2008 年，Milne 和 Witten[16] 介绍了利用维基百科丰富机器学习方法的方案；到目前为止，维基百科仍然是最有用的资源之一，无论是用于实体链接和消除歧义、语言建模、知识库还是其他各种任务。

2009 年，提出了远程监督 [17] 的概念。远程监督利用启发式或现有知识库中的信息生成带有噪声的模式，可用于从大型语料库中自动提取示例。远程监督现已被广泛应用，并且已经是关系提取、信息提取、情感分析等领域的常用技术。

特别指出近十几年来，随着大数据、神经网络、深度学习的快速发展，自然语言处理技术也大幅演化。下面也列举一些重要的里程碑。

2003 年，Bengio 等人提出基于神经网络的语言模型 [18]，用于替代传统基于计数和平滑的语言模型。语言建模是一种非监督学习形式，其被作为获得基础常识的先决条件。尽管语言建模很简单，但却是后续许多技术发展的核心。

2008 年，Collobert 和 Weston 将多任务学习 [19] 首次应用于 NLP 的神经网络。在他们的模型中，查询表（或单词嵌入矩阵）在两个接受不同任务训练的模型之间共享。多任务学习现在被广泛地用于 NLP 任务。充分利用现有的或人造的任务进行训练，可以更好的提高 NLP 效率。

2013 年，Mikolov 等人提出的 word2vec[20]，使语言模型的词嵌入训练极大加速，也使得大规模的词嵌入训练成为可能。

2013 年和 2014 年是开始引入神经网络模型到 NLP 任务的时期。使用最广泛的三种主要的神经网络是：卷积神经网络 [21][22]、循环神经网络 [23] 和递归神经网络 [24]。

2014 年，Sutskever 等人提出了序列到序列模型 [25]。这是一个使用神经网络将一个序列映射到另一个序列的通用框架。在该框架中，编码器神经网络逐符号处理一个句子，并将其压缩为一个向量表示；然后，一个解码器神经网络根据编码器状态逐符号输出预测值，并将之前预测的符号作为每一步的输入。机器翻译是对这个框架比较成功的应用。

2014 年，Bahdanau 等人提出注意力机制 [26]，是神经网络机器翻译的核心创新之一，也是使神经网络机器翻译模型胜过经典的基于短语的机器翻译系统的关键思想。序列到序列模型的主要瓶颈是需要将源序列的全部内容压缩为一个固定大小的向量。注意力机制通过允许解码器回

头查看源序列隐藏状态来缓解这一问题，然后将其加权平均作为额外输入提供给解码器。

大约同时在 2015 年前后，很多学者提出了基于记忆的网络。注意力机制可以看作是模糊记忆的一种形式。记忆由模型的隐藏状态组成，模型选择从记忆中检索内容。研究者们提出了许多具有更明确记忆的模型。这些模型有不同的变体，如神经图灵机 [27]（Graves 等，2014）、记忆网络 [28]（Weston 等，2015）和端到端记忆网络 [29]（Sukhbaatar 等，2015）、动态记忆网络 [30]（Kumar 等，2015）、神经微分计算机 [31]（Graves 等，2016）和循环实体网络 [32]（Henaff 等，2017）。

2018 年，Peters 等证明利用预训练语言模型 [33][34][35]，各种 NLP 任务都能获得大幅度提升改进。通过将语言模型嵌入作为特征，使用目标任务数对语言模型对进行微调 [36][37]，通常就能达到或超过传统结果。由于语言模型只需要无标记的数据便可以进行学习，因此对于标记数据稀缺的低资源场景，预训练语言模型尤其有用。

## 1.3 应用领域

自然语言处理研究的内容非常广泛，应用范围也非常广泛，如下举例一些常见的应用场景：

- 机器翻译：利用计算机实现自然语言（英语、汉语等）之间的自动翻译。
- 自动摘要：利用计算机自动地从原始文档中提取全面准确地反映该文档中心内容的简单连贯的短文。
- 文本分类：在预定义分类体系下，根据文本的特征，将给定文本于一个或多个类别相关联的过程。
- 情感分类：根据文本所表达的含义和情感将文本划分为褒扬或者贬义的两种或几种类型，是对作者倾向性、观点、态度的划分，因此也称倾向性分析。
- 信息抽取：从非结构化或半结构的自然语言文本中提取出于某个主题相关的实体、关系。事件等事实信息，并且形成结构化信息输出。
- 信息检索：用户输入一个表述需求信息的查询字段，系统回复一个包含所需要信息的文档列表。其核心技术在于索引构建和相关性计算。

- 问答系统：接受用户自然语言形式描述的问题，从大量异构数据中查找或者推断用户问题答案的信息检索系统。

# Chapter 2
# 语言模型与主题模型

本章简单介绍自然语言处理中的语言模型和主题模型的概念，并其常见的算法，以及在语音识别中的应用。

## 2.1 语言模型

### *2.1.1* 基本定义

语言模型（Language Model）用于计算语言序列 $w_1, w_2, \cdots, w_n$ 的概率，数学表示为 $P(w_1, w_2, ..., w_n)$，它是对语句的概率分布的建模。其最直接的应用就是判断一句话来自于人生成的语句的概率，例如在我们自然语言中，句子 “我去吃饭” 相比于 “吃饭去我” 的出现的概率更高，因此 $P(\text{“我去吃饭”}) > P(\text{“吃饭去我”})$。讲到这里，最直接的一个问题就是，如何计算 $P(w_1, w_2, ..., w_n)$ 呢? 我们下面介绍一种最基本的语言模型: n-gram 语言模型。

### *2.1.2 n-gram* 语言模型

n-gram 语言模型是一种最基础的语言模型。根据链式法则 (Chain Rule)，公式 $P(w_1, w_2, ..., w_n)$ 可以得到:

$$P(w_1, w_2, \cdots, w_n) = P(w_1)P(w_2|w_1)\cdots P(w_n|w_1, \cdots, w_{n-1})$$

其中的每一项 $P(w_i|w_1,\cdots,w_{i-1})$，可以用以下公式来估计，即:

$$P(w_i|w_1,\cdots,w_{i-1}) = \frac{C(w_1,\cdots,w_{i-1},w_i)}{C(w_1,\cdots,w_{i-1})}$$

其中，$C(\cdot)$ 表示该序列在训练语料中出现的次数。但是，当序列长度很长时候，计算 $P(w_i|w_1,\cdots,w_{i-1})$ 比较困难，一种常见的处理方式是引入马尔可夫假设 (Markov Assumption)，即假设当前词出现的概率只依赖于前 $n-1$ 个词，也就是:

$$P(w_i|w_1,\cdots,w_{i-1}) = P(w_i|w_{i-n+1},\cdots,w_{i-1})$$

根据 $n$ 取值的不同，我们可以得到不同的 n-gram 语言模型:

- Unigram: $P(w_1,\cdots,w_{i-n}) = \prod_{i=1}^{n} P(w_i)$
- Bigram: $P(w_1,\cdots,w_{i-n}) = \prod_{i=1}^{n} P(w_i|w_{i-1})$
- Trigram: $P(w_1,\cdots,w_{i-n}) = \prod_{i=1}^{n} P(w_i|w_{i-1},w_{i-2})$

### 2.1.3 n-gram 语言模型中的平滑技术

在计算 n-gram 时候，一个很重要的问题就是测试集中出现了训练集中未出现过的词而导致语言模型计算出的概率为零，我们称这些词为未登录词 (OOV)。平滑 (Smoothing) 技术就是为了缓解这类问题，常见的平滑技术有：拉普拉斯平滑 (Laplace Smoothing),、古德图灵法（good-turing)、线性减值法（Linear Discounting）等，感兴趣的读者可以深入阅读相关论文。

### 2.1.4 语言模型在语音识别中的应用

自动语音识别（Automatic Speech Recognition，ASR）是一种将人的语音转换为文本的技术，它是目前很多互联网产品如语音助手，语音搜索引擎等中必不可少的一部分。图 2-1给出了常见的语音识别系统的基本工作流程。其中基本可以分为以下几个模块:

- 数据预处理: 典型的预处理包含静音处理（Voice Activity Detection，VAD）等，用于去除其中的静音片段。

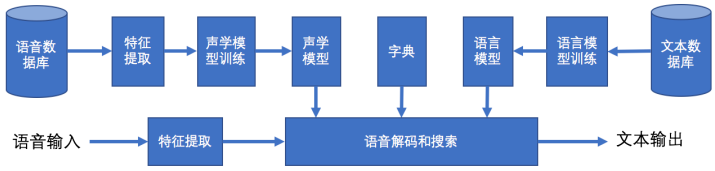



**图 2-1** 语音音识别基本流程

- 特征提取：将声音转换成包含声音信息的多维向量，常见的有 MFCC 等。
- 声学模型：主要是通过语音数据训练得到，其输出是音素等信息。
- 词典：字/词和音素之间的对应关系。
- 语言模型：也就是上文提到的语言模型部分，主要用于评估字或者词序列的概率。

语音系统首先将语音信号做特征提取工作，转化成诸如 MFCC 等特征来表示，然后使用语言模型和声学模型来解码，解码过程会产生很多候选 (Candidates)，最终最优的候选会被输出成为最终的结果。语言模型是其中很重要的一部分，它用于从根据语言统计规律评估声学模型给出的句子序列候选的概率，决定了最终输出的结果。

## 2.2 主题模型

### *2.2.1 基本概念*

主题模型 (Topic Models) 是近些年来非常重要的一项技术，它被广泛应用于工业和学术界。在主题模型中，我们一般用 $d$ 来表示要分析的文档，例如一篇文章或者一个网页等，而一个文档 $d$ 通常由一系列词 $(w_1, w_2, ..., w_n)$ 组成，其中 $w_n$ 是文档中的第 $n$ 个词。多份文档共同构成了我们要分析的语料集，我们用 $\mathscr{D}$ 来表示，$\mathscr{D} = (d_1, d_2, ..., d_m)$ 组成，其中 $d_m$ 是语料库中的第 $m$ 个文档。主题一般用 $z$ 来表示，它由一些词组成，同时也有该词在这个主题下的概率。主题模型泛指由一类可以从语料库

中抽取主题并利用这些主题表示文档的模型，常见的主题模型有 PLSA，LDA，以及各种 LDA 的变种，例如 SentenceLDA 等。在熟悉了这些基本概念之后，我们通过一种常见的主题模型 Latent Diriclet Allocation (LDA) 来认识主题模型。

### 2.2.2 常见的主题模型：*LDA*

2003 年 Blei 等人在《Latent Dirichlet Allocation》[13] 一文中提出了 LDA 模型。如图2-2所示，其中空心节点表示隐藏变量，实心变量表示客观测变量，整个模型具有 K 个主题，M 个文档和 N 个词。LDA 将文档的主题分布 $P(z|d)$ 看做随机变量 $\theta$，并且假设 $\theta$ 从一个狄利克雷先验中产生。同时，由于训练数据之外的文档对应的主题分布 $\theta$ 可以从上述狄利克雷分布中产生，训练数据之外的文档的 $\theta$ 可以更自然地进行计算。

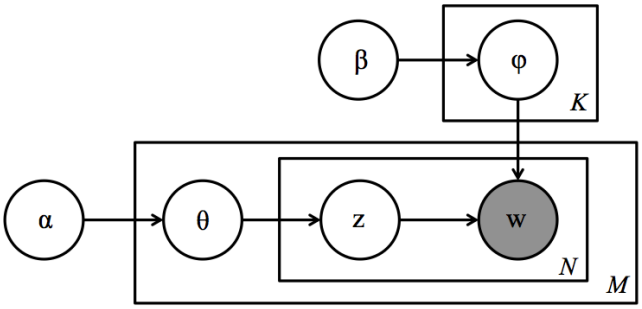



**图 2-2** LDA 图模型

### 2.2.3 主题模型在语音识别中的应用：语言模型适配

语音识别系统中一个常见的问题就是，我们训练语言模型的语料和它实际线上应用的语料之间存在不一致，这种情况下，除了重新训练模型，有一种代价更小的方法就是语言模型适配（Language Model Adaptation）。

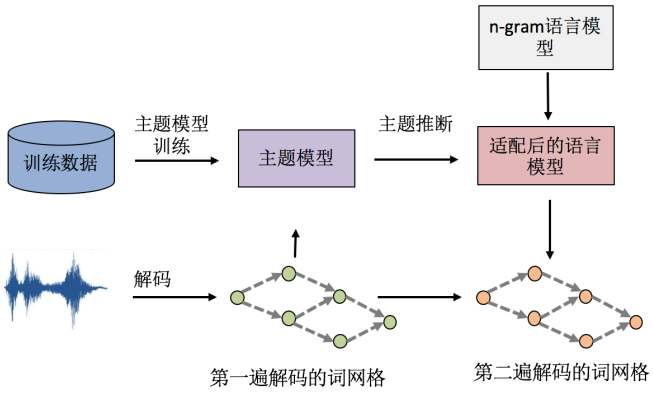



**图 2-3** 语言模型适配

语言模型适配指的是用实际应用的语料相关的信息，对语言模型做适配。图 2-3给出了其常见的工作流程，采用预先训练好的主题模型，我们对语音识别系统第一遍识别出来的词网格 (Word Lattice) 做主题推断，可以发现其语义级别的内容，同时它也可以作为一个 unigram 的语言模型 $P_{LDA}(w|\theta_d) = \sum_{k \in K} \varphi_{kw} \theta_{dk}$，对之前的 n-gram 语言模型就行适配：

$$P_d(w|C) = \lambda P_{LDA}(w|\theta_d) + (1-\lambda) P_{n-gram\ LM}(w|C) \tag{2.1}$$

其中 $C$ 代表当前词 $w$ 的上下文，$\lambda$ 是一个权重参数，$P_{n-gram\ LM}(w|C)$ 是基础的 ngram 语言模型给出来的评估分数。这个新适配过的语言模型，可以用于语音识别系统，重新解码出新的词网格以及最终的结果。

# Chapter 3
# 词法，语法与语义分析

## 3.1 词法分析

### *3.1.1* 问题定义

词法分析是自然语言处理的第一步，要做 NLP 深层次分析，比如句法分析、语义分析，甚至 NLP 复杂应用的先决条件，就是首先进行词法分析。词法分析的核心是，将自然语言解析为一个个词的序列，并判断每个词的词性、专名信息，为后续分析做好准备。总的来说，在中文这种孤立语中，词法分析主要由分词、词性标注、命名实体识别 3 个子任务组成。在英语、阿拉伯语等屈折语中，词法分析一般还包括词根还原（Word Stemming）任务。忽略词根还原，词法分析可以归纳为 3 个具体的子任务：

- 自动分词（Word Segmentation）：是将连续的自然语言文本，切分成具有语义合理性和完整性的词汇序列的过程。
- 词性标注（Part-of-Speech Tagging）：是指为自然语言文本中的每个词汇赋予一个词性的过程。
- 命名实体识别（Named Entity Recognition，也称专名识别）：是指识别自然语言文本中具有特定意义的实体，主要包括人名、地名、机构名、时间日期等。

如图3-1所示，将输入句子切分成一个个词汇，然后给每个词汇标记出名词、动词、介词等词性；并且识别出“2003 年 10 月 15 日”是一个时间专名，识别“杨利伟”是一个人名专名等。


**图 3-1** 词法分析示例

因为词法分析的自动分词、词性标注、专名识别本质上是类似的，所以它们的发展历史方也是类似的。总结起来，都大致经历词典匹配、机器学习、深度学习这 3 个发展阶段。

### *3.1.2 实现方案*

#### 3.1.2.1 词典匹配

基于词典匹配的实现步骤：

1. 词典构建：根据具体子任务的语言学知识构建词典。如果是自动分词，则收录常见的词条短语；如果是词性标注、专名识别，则收录当前常见词条对应的名词、动词、介词、专名类型等属性，并且保留相应词性、专名属性的概率。词典构建，通常需要人工收集、整理、离线更新，维护成本较高。
2. 词典匹配：扫描输入的所有子序列，如果当前子序列能够匹配词典中某个词条，则当前子序列属一个可能的候选。如果是分词，则当前子序列就可能为一个分词的词汇；如果是词性标注、专名识别，则当前子序列以相应的概率取词典中的词条属性。
3. 歧义消解：由于候选子序列之间存在歧义，所以基于词典匹配之上，需要加入一些启发式规则以解决歧义问题。常用规则主要包含最大前向匹配（Forward Maximum Matching，FMM）、最大后向匹配（Backward Maximum Matching，BMM）、最少切分、双向最大匹配、长片段优先等策略 [43][44]。

中文分词如图3-2所示。输入序列进行词典匹配之后，得到对应的 DAG 图，图中每条边都是词典词条，边的权重都为 1，图中每一种首尾

贯通的连接都是一种歧义切分的候选结果。比如由此只要求解 DAG 图的最优路径，则可以得到输入序列对应的分词结果。

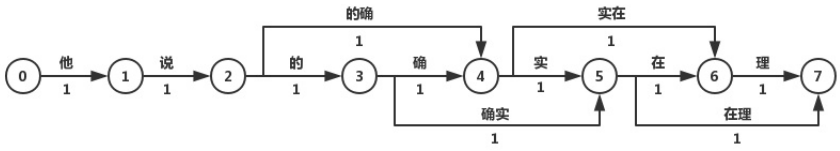



**图 3-2** 中文分词示例

很长一段时间内研究者都在对基于词典匹配方法进行优化，比如最大长度设定、字符串存储和查找方式以及对于词表的组织结构，比如采用 TRIE 索引树 [45]、哈希索引 [46] 等、AC 自动机 [47] 等结构方便快速查找。

基于词典匹配的优缺点：

- 优点：实现简单、运行速度快
- 缺点：面临词典收录更新困难、未登录词难处理；同时由于消除歧义策略过于简单，通常效果不甚理想。

#### 3.1.2.2 机器学习

基于机器学习的实现步骤：

1. 转换为序列标注任务：词法分析 3 个子任务，都通过定义标注空间标签集，将具体任务转换为标准序列标注任务。以中文分词为例：转换方法为，标注每个字在其所属词中的位置。因为对于任何一个字来说，它可以是一个词的开始（Begin），一个词的中间（Inside），一个词的结尾（End），或者本身就是一个单字的词（Singleton），这也就是在分词序列标注中常用的 BIES 的分类。只需将输入序列的每个字标上 BIES 标签中的一个，就可以转换得到对应的分词结果。这种标注空间（模型状态空间）的划分在词性标注和专名识别任务上也很常用，也会有一些类似的变种，比如专名识别中常用 BIO 标签集。

2. 求解序列标注任务：传统序列标注模型主要包括隐马尔科夫模型（Hidden Markov Model)[48][49][50][51]、最大熵马尔科夫模型（Maximum Entropy Markov Model）[52][53][54]、条件随机场（Conditional Random Field)[55][56][57]、结构化感知机（Structural Perception Machine）[58][59] 等浅层模型。这些浅层模型的区别主要在于如何对待输入字序列和标签序列之间的概率，训练目标是最大联合概率似然、最大条件概率似然，还是最小化风险等。

总结来说：传统序列标注模型中，CRF 是集大成者。相比于 HMM，CRF 去除了输出独立性要求，对于整个序列内部的信息和外部观测信息都可以有效利用，可以更加有效建模上下文。相比于 MEMM，CRF 通过全局归一化（global normalization），避免了 MEMM locally normalized 导致的 label bias 缺陷。

以自动分词任务为例，则其序列标注任务定义为：定义标签集为 $L = \{B, I, E, S\}$，给定输入文本序列 $X = \{x_1, x_2, ..., x_n\}$，目标是求解最优标注序列 $Y^* = \{y_1^*, y_2^*, ..., y_n^*\}$：

$$y^* = \arg\max_{Y \in L^n} p(\boldsymbol{Y}|\boldsymbol{X})$$

预测时，使用维特比算法求解最优标注序列 $Y^*$；训练时，使用最大化条件似然来训练模型，其中，$f_i$ 为人工定义的特征函数：

$$p_w(Y|X) = \frac{1}{Z_w(x)} \exp(\sum_i w_i f_i(y_i, y_{i-1}, x))$$

$$Z_w(x) = \sum_{y \in Y} \exp(\sum_i w_i f_i(y_i, y_{i-1}, x))$$

基于机器学习的优缺点：

- 优点：通过人工设计的特征工程，充分地挖掘了序列的上下文信息。模型的歧义消解胜过词典匹配的方法；同时具有很强的泛化能力，能够很好地处理未登录词问题。
- 缺点：特征工程不经需要耗费大量人力，而且需要大量语言学知识，设计和寻找有效特征存在较高门槛；另外，这些浅层模型，通常使用离散的 binary 特征，无法表达复杂先验，比如没法利用词向量。

#### 3.1.2.3 深度学习

随着大数据、神经网络、深度学习的快速发展，很多研究提出利用前馈神经网络来解决词法分析 [60][61] 任务。总的来说，与前面基于机器学习的方法类似，也是把词法分析任务作为序列标注问题进行求解，只是把人工设计特征函数，改成了使用多层前馈神经网络进行自动特征抽取。

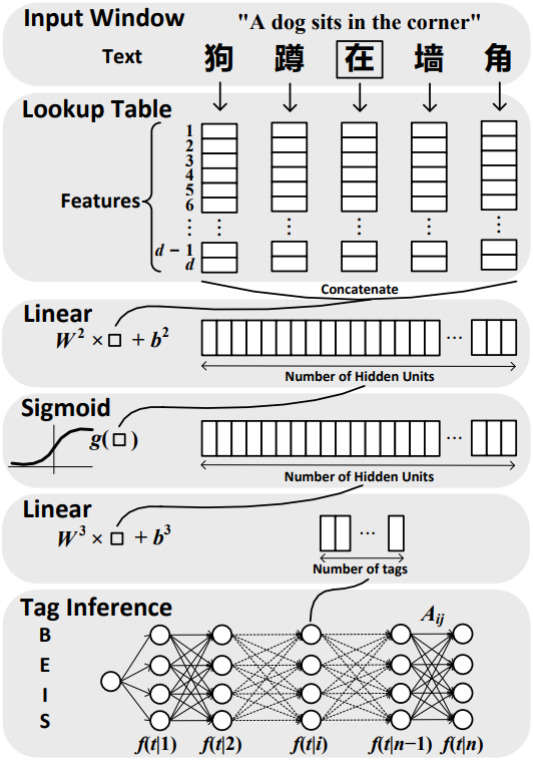



**图 3-3** 深度学习分词示例

如图3-3所示，网络的第一层输入句子中每个字的字向量，第二层将一个固定长度的字向量进行拼接，然后输入到标准前馈神经网络结构中，神经网络输出在标注集合上的 lattice，最后利用维特比算法进行解码，就可以得到最优标注序列。

后续有很多研究，深度学习框架基础上，对前馈神经网络进行改进，进一步探索了卷积神经网络 [62]、循环神经网络 [63]、递归神经网络 [64] 等复杂结构对词法分析任务的影响。

另外，传统的词法分析通常会把分词、词性标注、命名实体识别当成 pipeline 形式的进行处理，这样带来的一个问题就是错误传播。比如，分词的错误将会导致后续词性标注、专名识别、句法分析、语义分析出现错误。所以在学术界也有很多联合建模 [59][65][66] 方面的工作。联合建模的一大好处是词法分析与其他任务可以共享有用的信息，词法分析的时候也会考虑到其他任务的要求，其他任务也会考虑各种词法分析的可能性，通常可以再全局上取得最优解。但是随之而来问题是搜索的复杂度往往会显著提高：需要更有效的搜索剪枝机制在控制复杂度的同时，不对搜索的结果产生显著影响。

值得一提的是，2018 年 Peters 等证明在利用预训练语言模型 [33][34][35]，各种 NLP 任务都能获得大幅度提升。将语言模型嵌入作为特征，使用目标任务数对语言模型对进行微调 [36][37]，通常就能达到或超过传统结果。由于语言模型只需要无标记的数据便可以进行学习，因此对于标记数据稀缺的低资源场景，预训练语言模型尤其有用。

总结起来，基于深度学习方法的优点是通过深度神经网络自动学习多层特征抽象，避免了复杂的特征工程；模型的歧义消解、泛化能力通常都很好。只是，深度学习模型需要训练充分，通常对数据量、计算量都有较高的要求。

### *3.1.3 应用案例*

通常 NLP 中深层次的语法语义分析通常都是以词作为基本单位，所以词法分析都是这些深层次分析的基础和先决条件。词法分析作为自然语言处理的第一步，是下游所有分析任务的基础。这些下游应用小句法分析、语义分析，大到对话系统、文本分类、自动摘要、机器翻译 [67]、信息检索、搜索引擎、语音合成等等。几乎只要有利用到自然语言处理技术的地方，词法分析都是不可或缺的基础技术。

比如在搜索引擎中，用户输入一个表述需求信息的查询字段，系统回复一个包含所需要信息的文档列表。其核心技术在于索引构建和相关性计算。

索引构建：首先需要对互联网海量文档进行，分词、词性标注、专名识别；利用词法分析的结果，以词为粒度，为所有文档建立倒排索引。

相关性计算：首先需要对用于查询字段进行词法分析，找出表达用户需求的核心词汇；然后以这些核心词汇为 key，去拉取相应的倒排索引，获取初步筛选后的文档；接着，为了要提高返回结果的相关性，需要计算用户查询和初筛文档之间的相似度，取相似度较好的文档返回。其中计算用户查询和文档之间的相似度，通常利用 BM25、语义计算等技术，而它们也都是以词为单位进行的，由此看出词法分析在搜索引擎中的重要性、基础性。

在比如在文本分类和情感分类中，在预定义分类体系下，输入给定文本，抽取文本特征，将给定文本于一个或多个类别相关联的过程。其中，抽取文本特征的时候，通常也都是先进行词法分析，先获得给定文本的词汇、词性、专名等信息，然后以词汇为单位，以词汇的词向量、词性专名等信息为特征，利用机器学习、深度学习的分类模型进行预测。

## 3.2 语法分析

### *3.2.1* 问题定义

短语结构语法（Constituency Structure Grammar）和依存关系语法（Dependency Grammar）是现在常见的两种语法关系。短语结构语法又叫上下文无关文法 (Context-Free Grammars，CFGs)，它从一个特殊的初始符号出发，不断的应用一些产生式规则，从而生成出一个字串的集合（如句子）。产生式规则指定了某些符号组合如何被另外一些符号组合替换。它呈现一个树分类关系，句法根据一定的规则进行转换分析。每一个词的转换都是需要按照设定的树值规则进行目的性的转换。

依存语法 (从属关系语法) 是由法国的语言家 Lucien Tesnière 提出的 [68]，它将句子各个词语之前的搭配关系描述成预先定义好的依存关系。它基于一个基本假设：句法结构本质上包含词和词之间的依存（修

饰）关系。一个依存关系连接两个词，分别是核心词（head）和依存词（dependent）。依存关系可以细分为不同的类型，表示两个词之间的具体句法关系。比如主语依赖于谓语（SBV），宾语也依赖于谓语（VOB）以及定语依赖于名词性短语（ATT）等。依存句法认为“谓语”中的动词是一个句子的中心，其他成分与动词直接或间接地产生联系。

除了以上介绍的两种句法体系外，国内外都开展了对句法分析的研究。不论是国外的链语法 (Link Grammar)、组合范畴语法（Combinatory categorial grammar，CCG）等，还是国内黄曾阳教授提出的 HNC 理论 (Hierarchical Network of Concept)[69] 都是目前行业内常用的语法，只是由于设定区域的不同，所以使用有一定的局限性。

短语结构分析的语法集是由固定的语法集组成，较为固定和呆板，依存语法则更加的自由。另外依存语法树标注简单且 parser 准确率更高，再加之通用依存数据集（Universal Dependencies Treebanks）的发展，依存语法分析受到专家学者普遍的青睐，得到越来越多的应用。这里也将着重介绍依存语法分析。


**图 3-4** LTP 依存句法分析样例

例如，从上述例子中我们可以看到，句子的核心谓词为“强调”，主语（SBV）是“他”，强调的宾语（VOB）是“紧迫性”，“紧迫性”的修饰语（ATT）是“武汉疫情控制”。

### *3.2.2 实现方案*

依存句法分析方法主要可以分为两种，一种是基于图的方法（Graph Based），一种是基于转移（Transition Based）的方法。基于图的方法先

建立句子中所有词语的全连接图，然后求图中的最大生成树。两种方法中更主流的算法是基于转换的依存句法分析，基于转移的方法将依存树的构成过程建模为一个动作序列，将依存分析问题转化为寻找最优动作序列的问题。通过 SHIFT，LEFT_ARC，RIGHT_ARC 三个动作来将序列转换为树结构。一次分析任务 $C=(s,b,A)$ 由一个 Stack 栈 $s$，一个 buffer 缓冲区 $b$，一系列依存弧列表 $A$ 构成。初始化栈 $s$ 里面只包含一个 *root* 元素即根元素，$s_1$ 代表栈顶元素，$s_2$ 表示栈顶第二个元素。缓冲区 $b$ 是一个队列，里面包含了要解析的一句话的序列，$A$ 为空。一条依存弧有两个信息：动作类型 + 依存关系名称 $l$。$l$ 视依存句法语料库中使用了哪些依存关系 label 而定。动作 SHIFT 将缓冲区 $b$ 中最上面的一个词 $b_1$ 移到 stack 中，即不建立依存关系，只转移句法分析的焦点，即新的左焦点词是原来的右焦点词，依此类推。LEFT_ARC，即添加栈顶两个词 $s_1$，$s_2$ 之间的依存边，方向为 $s_1 \rightarrow s_2$，并且将 $s_2$ 从栈 $s$ 中删除。RIGHT_ARC，即添加栈顶两个词 $s_1$，$s_2$ 之间的依存边，方向为 $s_2 \rightarrow s_1$，并且将 $s_1$ 从栈 $s$ 中删除。

比如图3-4中，首先模型判断执行 SHIFT 动作将“他”移到栈 $s$ 上，然后再执行 SHIFT 动作将“强调”移到栈 $s$ 上，接着模型判断这两个词之间有依赖关系，且方向为为“强调 $\rightarrow$ 他”，执行 RIGHT_ARC 动作在 $A$ 中添加一条依存弧，且将“他”从栈 $s$ 中删除。接着模型判定执行 SHIFT 动作将“了”移到盏 $s$ 上，然后模型判断执行 LEFT_ARC 操作，即添加“强调”和“了”之间的依存弧“强调 $\rightarrow$ 了”到 $A$ 中且将“了”移出栈 $s$。如此往复，直到最后当缓冲区 $b$ 为空和栈 $s$ 只有 *root* 时结束。训练模型的主要目标即寻找一个分类器，当给定一个 Configuration (当前的 stack，buffer，依存弧列表）时预测下一步转移的操作类别。

基于转移的解析过程是线性的，动作步骤随句子长度线性增长，而基于图的方法需要在全图上做搜索，所以时间复杂度上基于转移的方法会有优势。但是基于转移的方法在解析的每一步都只是利用局部信息，会导致错误传播，性能比基于图的略差。

近几年，分别出现了针对这两种不同方法的神经网络模型。比如基于图的 [70, 71, 72]，直接用神经网络来预测每两个词之间存在依存关系的概率，得到一个全连接图，图上每个边代表了节点 $a$ 指向节点 $b$ 的概率，然后使用 MST 等方法来来将图转换为一棵树。概率的计算可以简单的使用节点 $a$ 和节点 $b$ 的 embedding 向量做向量运算，也可以使用复杂的多

层 GNN 网络迭代更新。基于转移的如 [73, 74, 75]，通过两个 LSTM 来分别建模 stack 状态、buffer 状态，使用第三个 LSTM 网络或者 Pointer 网络来建模动作序列。

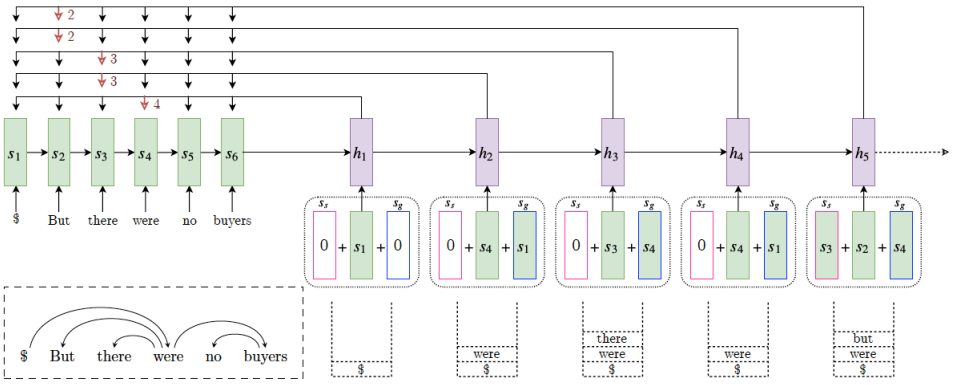



**图 3-5** 基于转移的算法 Stack-Pointer Networks for Dependency Parsing 网络结构图

### *3.2.3 应用案例*

依存语法分析在信息检索，评价抽取和情感分析等 NLP 任务上都有很多应用。比如“谢霆锋的儿子是谁”和“谢霆锋是谁的儿子”两句话如果不使用依存语法分析，很有可能就返回了一样的结果。依存句法分析能够更直接地通过语法结构的规则约束筛选出可能正确结果，提升相应任务的准确性。又比如“深圳大学非常漂亮，学生都很聪明”，这里“漂亮”形容的是“深圳大学”，“聪明”形容的是“学生”，通过依存句法分析，就可以抽取出对应的搭配。再比如“我家音响声音很大”和“我家洗衣机声音很大”，两者在情感上前者是正评价，后者是负评价，需要使用依存句法分析来识别“声音很大”的修饰对象。

常用的中文依存句法分析的工具有复旦大学 fnlp[1]，斯坦福大学 Stanford CoreNLP[2]，Hanlp[3]和哈工大 LTP[4]。

## 3.3 语义分析

在自然语言处理领域，语义分析涉及在某种程度上理解单词、短语、句子或文档的意义。传统狭义语义分析主要包括语义消歧（word sense disambiguation）[76] 和语义角色标注（semantic role labeling）[77, 78]。语义消歧指在给定文本上下文中确定多义词语的含义，例如，“他买了一台新苹果，用来修图更方便了”中，苹果一次指代苹果电脑，而非水果。语义角色标注是给词语和短语标注其在上下文文本中的含义的过程，典型标注标签包括主体、意图、结果等，如图3-6。语义角色标注是面向任务型对话系统中核心组件自然语言解析模块的基础技术之一。

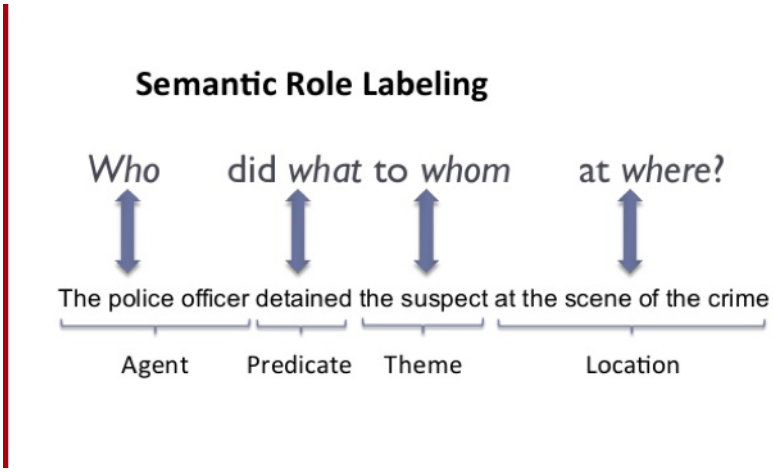



**图 3-6** 语义角色标注示例

随着深度神经网络在自然语言处理领域的应用和研究，基于神经网络的语义分析得到了越来越多的应用和发展，并成为了驱动神经机器翻

[1] https://github.com/FudanNLP/fnlp
[2] https://stanfordnlp.github.io/CoreNLP/
[3] https://github.com/hankcs/HanLP
[4] https://github.com/HIT-SCIR/ltp

译 [26, 79]、阅读理解 [80]、对话系统 [81, 82] 的最基础和核心的技术。基于神经网络的语义分析广义上可分为语义表示和语义匹配。其中语义表示任务将词语和短语镶嵌到高维向量空间中，称作词向量，作为 CNN[83]、RNN[84] 和 Transformer[85] 等模型的底层输入，根据任务需要，可以灵活的实现文本分类 [86, 87]、机器翻译、阅读理解、对话系统、文本摘要 [88, 89] 等自然语言处理任务，逐步替代或部分替代了传统以 ngram one-hot 向量作为底层表示的方法。语义匹配任务更多关注句子和篇章层级的语义的相似性，例如，寻找给定语料库中与查询语句语义层面最相似的句子。语义匹配任务在信息抽取、对话系统、问答系统都有广泛的应用。

### *3.3.1* 语义表示

语义表示中词向量的概念可以追溯到 Bengio 的著作 [18]。作者在文中提出了一种基于神经网络的语言模型，创新的提出将每一个词表示成一个高维的向量，而后使用神经网络计算给定上文词语后，下一个词出现的概率分布，最大化训练语料出现的概率。训练语料是单词序列 $w_1, \cdots, w_t, w_t \in V$，其中 $V$ 是全体单词的集合。语言模型的目标是找到一个模型 $f(w_t, \cdots, w_{t-n+1}) = P(w_t | w_1^{t-1})$，即在给定前 $t-1$ 个词的情况下，对第 $t$ 个词的概率分布建模。Bengio 提出把词表示成实数向量，$C(i) \in \mathbb{R}^m$，即每一个词对应一个维度为 $m$ 的实数向量，映射 $C$ 可以用一个 $|V| \times m$ 的实数矩阵表示。在第二步中，作者提出使用神经网络 $g$ 来建模给定前缀序列下一个词的概率。模型结构如图 3-7。其中，将词表示成高维实数向量的方法是词向量以及一系列神经网络在自然语言处理应用的开端。通常，我们会设计一个仅与文本结构或上下文相关的（无监督）任务作为目标，拟合训练数据，得到词向量表示。在 Bengio 之后，Word2Vec[90], Glove[91]，ELMo[92] 等一系列词向量方法出现并逐渐成为自然语言处理的标准工具。语义表示不仅限于词的层面，广义的语义表示在现代基于神经网络的自然语言处理中应用广泛。例如，神经机器翻译通常采用编码器译码器（Encoder-Decoder）[25] 结构，其中编码器的输出可以看做是翻译模型对输入文本的语义表示，译码器基于该表示，生成目标语言的文本。再如，闲聊型对话系统中，需要针对聊天的上文给出合适的回答。闲

聊对话系统可以分为生成式对话系统（Generation-based）[93] 和选择式对话系统（Selection-based）[82]。选择式对话系统依赖语义匹配方法。典型的生成式闲聊对话系统也采用编码器译码器结构，其中编码器的输出是对聊天上文的总结性表示，可以认为是广义上的聊天上文的语义表示，用于在译码器中生成合适的回答语句。此外，随着自然语言处理预训练模型的发展，Bert[94]，GPT-2[95]，MASS[96] 等预训练模型（Pretrained models）在阅读理解等任务上大放异彩。这些模型可以看做是多个联合任务共享底层语义表示的学习，通常使用与文本结构相关的目标作为训练目标，设计神经网络结构表示文本。在应用时，这些预训练获得的表示仅需在少量标注数据上进行优化，即可获得优秀的子任务模型。通常，编码器和预训练模型采用 CNN、RNN、Transformer 等结构处理文本序列，在 CNN 和 RNN 中通常还会使用注意力（Attention）机制，得到文本片段的语义表示。

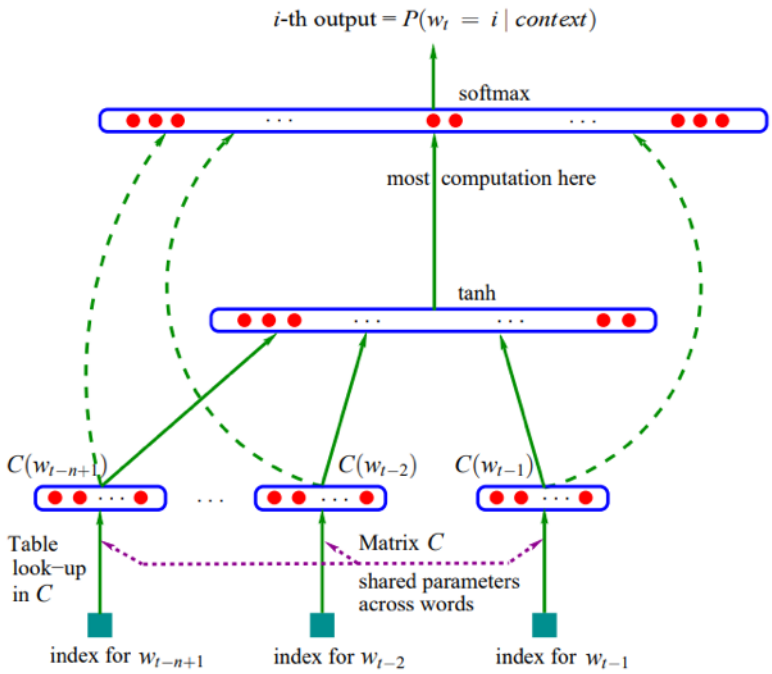



**图 3-7** 一种使用词向量的神经语言模型结构

### 3.3.2 语义匹配

语义匹配任务在语义表示的基础上，对文本片段之间的相似度或相关度给出量化指标，语义越相似的片段匹配分数越高。语义匹配可大致分为基于表示的匹配（representation-based matching）[97] 和基于交互的匹配（interaction-based matching）[98, 99]。基于表示的匹配方法注重对表示层的构建。常见的匹配度计算方式包括 cosine 函数，和神经网络匹配，如图 3-8。Cosine 函数直接计算待匹配对的语义表示向量间的 cosine 值，通常 cosine 值越大代表待匹配对的相似度或相关度越高。这种方法不需要额外的训练数据，实现简单并且高效，在工业中广泛应用。神经网络匹配方式使用一个额外的神经网络结构，将待匹配对的语义表示向量作为输入，计算语义匹配分数。这种方法更加灵活，可根据数据定制匹配结构。但需要额外的标注数据进行训练，才能得到可用的匹配模型。


**图 3-8** 基于表示的匹配

基于交互的匹配方法通常会保留待匹配对的序列信息，不会在表示层将文本转换成唯一的一个整体向量表示，而保留一个向量序列，用于接下来的交互匹配过程。例如将句子表示成一个与句子等长的向量序列。该向量序列可以使用语义表示的方法得到，例如使用 RNN 对文本序列进行建模。得到待匹配的序列对后，可以对序列中每个位置的向量计算与待匹配序列中向量的相关度。对每一个位置使用相同的方式计算相关度，可以得到一个匹配矩阵（matching matrix）。匹配矩阵包含了更细致的局部文本交互信息，在交互矩阵的寄出上，我们可以构建神经网络结构计算最终匹配得分，去拟合目标得分。图 3-9展示了一个典型的基于交互的语义匹配方法在选择式对话系统中的结构。其中 Representation module 对应语义表示部分，matching block 对应交互匹配部分。

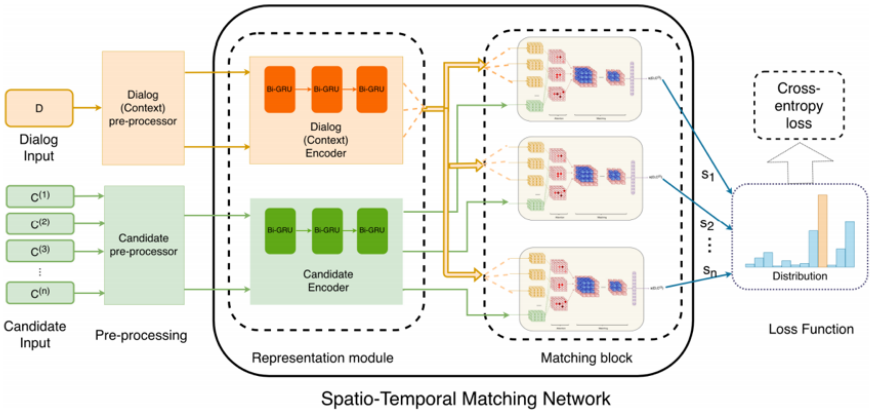

D
Dialog Input
Dialog (Context) pre-processor
Bi-GRU
Bi-GRU
Bi-GRU
Dialog (Context) Encoder
C(1)
C(2)
C(3)
C(n)
Candidate Input
Candidate pre-processor
Pre-processing
Candidate Encoder
Representation module
Matching block
Cross-entropy loss
s1
s2
sn
Distribution
Loss Function
Spatio-Temporal Matching Network

**图 3-9** 基于交互的匹配

# Chapter 4
# 对话理解与智能质检

本节首先给出对话理解任务的定义，然后介绍对话理解的主要方法。接下来以智能质检为例，讲述对话理解是怎么落地和实现的。

## 4.1 对话理解

### *4.1.1* 什么是对话理解

对话理解是指希望机器像人一样，具备语言理解能力，从对话内容中挖掘意图，理解意图，情绪识别等。例如，在客服与用户交互的对话中，用户询问“今天的天气如何”，这里就是一个“询问天气”的意图。在实际对话中，由于语言的多义同义问题，语言的词序问题，我们不能只停留在字面理解层面，更需要语义层面的理解。

这里对话可以包含语音对话和文本对话，如果是语音对话，我们一般可以利用语音自动识别技术将语音转为文本。后续我们要讨论的内容是文本的对话理解。

### *4.1.2* 技术路线简介

一般而言，文本的对话理解从技术角度上可以分为两类：文本匹配和文本分类。

**文本匹配** 文本匹配的目标是得到 $f(text_1, text_2)$ 的语义匹配得分，其中 $text_1$ 和 $text_2$ 是输入的文本，$f$ 是文本匹配模型。传统的文本匹配技术如信息检索中的向量空间模型 VSM、BM25 等算法，主要解决词汇层面的匹配问题。实际上，基于词汇重合度的匹配算法有很大的局限性。例如“出租车”和“的士”语义是一致的，但是字面上却完全不匹配。近年来，随着深度学习技术的发展，我们通过多层神经网络对文本语义进行建模，在语义匹配效果上有了很大的提升。

深度学习的模型主要包含 Representation-based Match 和 Interaction-based Match 两种，如下图所示。

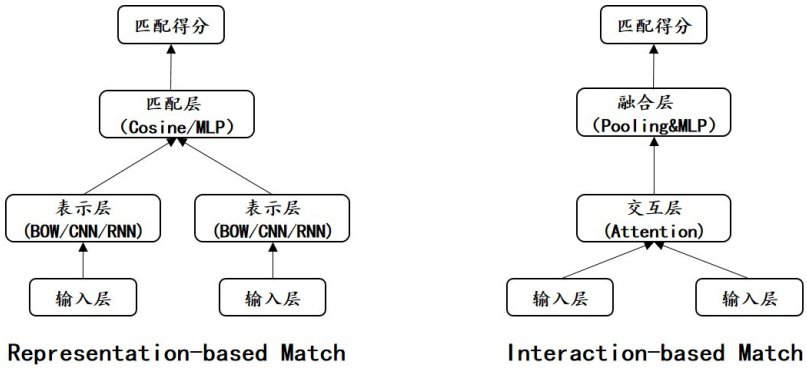



**图 4-1** Representation-based Match 和 Interaction-based Match 框架图

输入层一般是将词通过 look-up table 变成词向量，常见的方法有 word2vec, 其后陆续出现了一些文本向量化方法, 如 GloVe, FastText, EMLo 等。但在相当多的应用场景中, word2vec 几乎仍是最适合的方法之一。词到向量的转化使得词与词之间的语义相似性可以通过向量相似度的方法来度量。为了减少数据稀疏，我们可以加入 subword information，即词内的结构信息，中文里面就是字的信息，英文就是字母的信息。

Representation-based Match 的核心思想是先将句子转为向量，然后再匹配。怎么获得句子向量呢？最简单的方式有 Bag of Words（BOW），句子中词向量求和作为句子向量表示，相加还有一个作用是将变长的向量转化为定长。BOW 的方式忽略了句子中词序对语义带来的影响。当然，句子向量更高级的获取方式可以使用卷积神经网络（CNN）或者循环神经网络（RNN）来得到。经典的方法有 Siamese CNN，Siamese LSTM，

DSSM，ARC-I 等。Representation-based Match 的优势是可以离线计算好句子向量的表示，易于建立索引，执行效率高，非常适合文本匹配的粗召回。但其存在的问题是容易失去语义焦点，词的上下文重要性较难衡量。

Interaction-based Match 则是构造了两个句子之间的语义单元（e.g., term，n-gram，part-of-speech）的交互矩阵，然后再经过一个融合层将细粒度的语义匹配信息做融合，较好的把握语义焦点和保留重要的句子间相似信息。主流的方法包含 ARC-II，DeepMatch 和 MatchPyramid 等。Interaction-based Match 不能离线预处理，需要在线匹配，适合精排序。

**文本分类** 文本分类的目标则是得到 $g(text)$ 的语义标签，其中 $text$ 是输入的文本，$g$ 是文本分类模型。文本分类问题需要数据标注其语义标签，在任务型对话中，这个语义标签就是意图。文本分类模型训练好后，我们就可以对新数据进行分类了。具体内容可以参考对话系统的章节。

## 4.2 应用案例：智能质检

对话理解在智能客服，智能质检有着广泛的应用。下面以智能质检为例，阐述对话理解相关技术是怎么应用的。

### *4.2.1* 什么是智能质检

智能质检使用自然语言算法和机器学习算法，分析呼叫中心场景下客服坐席人员与客户的对话，实现全量质量检查，提高坐席效率和客户满意度。智能质检系统的输入是一通人工客服和客户对话的录音，输出是质检报表，显示该录音在不同质检项的合格情况。质检项的重要性通过质检项的分数来决定。智能质检无需人工介入，节省质检人力，覆盖率高（100%），提升质检效率，降低漏检错检率。

从智能质检的基本流程图中我们可以发现，对话录音经过语音识别模块之后，我们得到了客户和人工客服之间的对话文本。质检员配置了质检项之后，我们将对话文本输入质检模型，最后得到了质检报表。

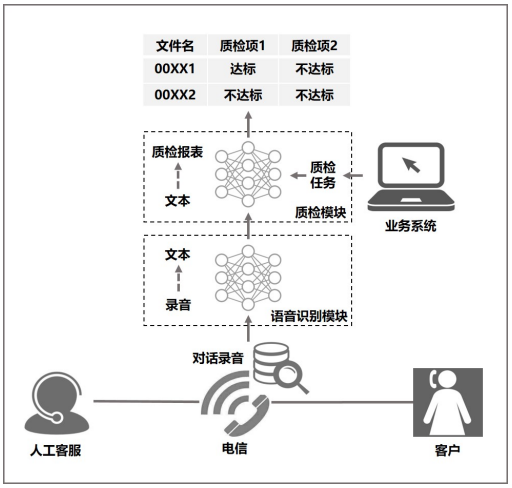



**图 4-2** 智能质检基本流程图

### *4.2.2 实现方案与应用状况*

**实现方案** 假设我们定义了质检项要求客服在对话中“核实用户的工作地址”，比如“你的公司地址在哪里”。一种简单的实现方案是通过规则配置“算子 + 逻辑操作符”或者正则表达式，如果录音文本中满足匹配条件，则命中该质检项。我们可以配置“(公司 | 工作)& 地址 &(哪 | 地方)”这样识别出“你的公司地址在哪里”和“您的工作地址在什么地方”这两种说法。这里“公司”就是一个关键词算子，“&”是一个“与”的逻辑操作符，“|”是一个“或”的逻辑操作符。这种本质上还是一种基于词汇重合度的匹配算法，适合冷启动，在没有标注数据的情况下就能构建一个简单的质检系统。但它也存在诸多弊端：第一、配置规则需要一定的专业知识，有培训成本；第二、当质检项和表达方式增多时，不同质检项容易出现规则冲突，维护成本高；第三，对于一些比较复杂的质检项，很难通过配置规则进行质检，容易出现漏检和误检的问题。

如果我们有标注数据，就可以使用文本匹配和文本分类的方法。计算 $f$(录音文本, 质检例句) 的语义匹配得分或者 $g$(录音文本) 的质检项标签。为了减少对数据的依赖和利用大量的无标签数据，一种有效的做法是在 BERT（Bidirectional Encoder Representation from Transformers）模型上去 fine-tune 分类器，BERT 用于上层模型的特征提取，作为上层模型

的输入，它能够较好的捕捉对话片段的高层语义，兼容少量的漏字情况。我们通过数据驱动的方法让模型越来越聪明，业务方只需要提供标注数据就能进行质检，不需要人工定义规则，模型具有一定的泛化能力。但遇到 bad case 没有基于规则的方法容易修复，另外需要标注数据积累到一定规模才能发挥模型的优势。

**应用状况** 当前智能质检的应用可以包含离线质检和坐席实时质检。离线质检是指结合语音识别和自然语言处理技术，对海量录音数据进行批量的智能化分析。离线质检在质检过程无需人工介入，可以提供内容质检，敏感词识别等质检结果。坐席实时质检是指在人工客服和客户通话过程中，提供实时质检功能，辅助人工客服判断客户情绪和实时分析对话过程的信息，及时提醒人工客服从而使客户获得更好的服务。现在智能质检的产品形态包含 SaaS 云服务和私有化部署。SaaS 级产品部署，让中小企业也能够享受智能质检带来的高效与便捷，克服了采购费用高部署周期长的问题。

未来随着多方业务的使用，可以基于联邦学习进行智能质检，在满足数据安全和私隐保护的前提下，通过模型的参数梯度共享，获得了把所有数据放在一起训练的效果，使得不同的业务方合作共赢，建立更准确的数据模型。

# Chapter 5
# 人脸识别

## 5.1 问题定义

人脸识别，是指对输入的图像和视频，检测其中存在的人脸，依据人脸的面部特征，完成身份识别的过程，属于生物特征识别技术。整个流程包含人脸检测、人脸对齐、人脸特征提取、人脸匹配几个阶段，如图 5-1 所示。目前人脸识别已经广泛应用于安防、金融、军事等领域。

人脸识别具有以下优点：

自然性：所谓自然性，是指人脸识别技术所利用的生物特征，与人类进行人脸识别时所利用的生物特征是一致的，与之相比，虹膜识别、指纹识别等技术，则不具备自然性。

非接触性：在人脸识别技术中，用户不会与识别设备发生任何接触，对于用户来说体验较好。而指纹识别则需要用户进行按压设备。

使用便捷：用户使用人脸识别技术时非常方便，基本上无需做特殊的配合。

人脸识别也具有一些缺点，比如易受光照条件的影响，易受人脸遮挡物的影响，跨年龄识别难度较高等。但总的来说，人脸识别是目前一种可靠的，实用的，便捷的身份核验技术。


**图 5-1** 人脸识别流程图

## 5.2 实现方案

### *5.2.1* 人脸检测及对齐

人脸检测是人脸识别的第一步，属于目标检测的子方向。其目的是找出图像中的人脸以及对应的位置。可能还会包含一些人脸的额外信息，比如人脸的关键点，姿态角度等。

典型的人脸检测是基于以下的流程：由于人脸可能出现的图像的任何位置，因此需要通过滑动窗口（sliding windows）来获取可能包含人脸的子图像。获取到的子图像，需要通过一个二分类的分类器，来判断图像中是否包含人脸，如果还需要确定人脸的精确位置，还需要加上一个回归人脸框的操作。同一个人脸可能会检测出多个人脸框，因此需要使用非极大值抑制 (Non-Maximum suppression, NMS) 来进行合并去重。由于人脸有大有小，为了更好的检测出不同尺度的人脸，还需要输入图像做不同尺度的图像缩放，也叫图像金字塔技术。接下来本文介绍一些具有代表性的人脸检测方法。

Viola-jones[100] 是 2001 年提出的一个人脸检测方法，该方法具有检测效率高，并且能够保持较好的精度的特点，是第一个具有实用意义的人脸检测算法。该方法使用 Haar-like 小波特征，并通过级联的 AdaBoost 分类器构造检测器。同时，还是用了积分图来加速 Haar-like 输入特征的计算。这是一个具有里程碑意义的方法。

MTCNN[101] 是 2016 年提出的一个人脸检测方法，它将人脸分类、人脸框回归以及人脸关键点定位在同一个任务内完成了，是一个多任务 (multi-task) 的检测方法，这种思路在后续的很多方法里也得到了使用。该方法采用三级网络级联的方法，定义了 3 个子网络：P-Net,R-Net 以及 O-Net,P-Net 用于快速产生候选框，使用全卷积运算代替了滑动窗口，大大提升效率，R-Net 和 O-Net 用于对候选框进行精细调整。

anchor 的思想在 Faster-rcnn[102] 中首先被提出，该方法被广泛应用于两阶段和单阶段的目标检测任务中，在人脸检测中也经常使用到。anchor 提出的目的是为了解决目标在图像中可能以不同的形状存在，比如不同的长宽比，所以加入人工的先验信息，预先定义不同比例的 anchor 来进行候选目标框的获取。Face r-cnn[103], Pyramidbox[104], Retinaface[105] 这些方法，都用到了 anchor 的思想。另一个在人脸检测中经常使用的思

想是特征金字塔网络 (feature pyramid network)，人脸在图像中，可能以不同尺度的大小出现，为了能够对大人脸和小人脸都有很好的检测效果，一般有两种做法，一种做法是图像金字塔，这种方法需要对输入图像做不同尺度的缩放，缺点是耗时较高；另一种更好的做法则是特征金字塔，其思想是在不同分辨率的特征图 (feature map) 上检测对应尺度的目标，同时将不同分辨率的特征图与更高层的特征图进行特征融合，保证每一层的特征图都具有足够的表达能力。Pyramidbox,Retinaface 都用到了特征金字塔，SSH[106] 虽然没有直接用到特征金字塔，但其也是对网络 3 个不同尺度的特征图进行分别预测，来解决多尺度的人脸检测问题。

人脸检测检测通常会以检测率、误检率等作为评估指标，同时会关注性能指标。很多算法的效果很好，但性能会很差，做不到实时，也就限制了使用场景。

做完人脸检测后，一般需要进行人脸对齐。通过对人脸进行关键点定位，以及预先定义好的关键点模板，进行仿射变换，通过旋转、平移、缩放等操作，进行人脸对齐，对齐后的人脸能够更好的进行人脸特征提取。目前常见的关键点个数，有 5 个关键点、68 个关键点、90 个关键点以及 106 个关键点等。

### *5.2.2* 人脸特征提取

人脸识别中，最重要的步骤是人脸特征提取，将人脸在高维空间的描述转化为其他空间内的低维描述，使得转化后的特征能够很好地区分不同人之间的差异点。经过特征提取得到人脸的特征表示之后，可以进行特征匹配。如果是对两个特征进行比对，我们一般称为人脸比对或者人脸验证 (verification)，如果是将一个特征与一组特征进行匹配，我们一般称为人脸检索或者人脸识别 (identification)，如图 5-2 所示。匹配的效果直接依赖于特征提取是否准确，因此如何提取好的人脸特征是特别重要的。

传统的特征提取算法，通过一些降维方法，得到一系列降维后的特征，用来表示人脸。比如使用 PCA 进行降维的 EigenFace，基于 LDA 进行降维的 FisherFace 等，都是早期人脸识别中非常经典的算法。但这些方法存在一些缺点，对光照、表情、姿态敏感，泛化能力不足，因此在实际使用中的准确度不高。

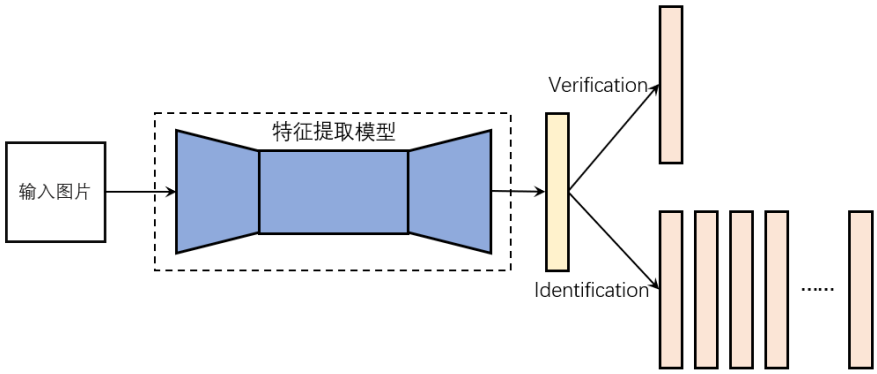



**图 5-2** 人脸特征抽取

随着深度学习的广泛应用，越来越多具有实用价值的方法被提出，人脸识别的研究得到了极大的发展。基于深度学习的特征提取方法可以分为两大类：

度量学习 (metric learning)

通过一个度量函数，来衡量相同人或者不同人的特征表示之间的距离，从而学习到每个人每个照片的特征表示，基本思路是同一个人的特征表示之间的距离尽可能小，不同人的特征表示的距离尽可能大。这一个方向的典型方法包含 2014 年的 Contrastive loss[107] 和 2015 年 google 提出的 Triplet loss[108]。

DeepID2 是基于 Contrastive loss 的模型，它在训练的时候，同时训练 classification 和 verification 两个信号，其中的 verification 信号，就是用特征表示之间的 Contrastive loss 来构造的。Contrastive loss 是基于 pairwise 的思想，模型训练时，需要输入两张图片，如果两个图片是同一个人，则 verification 的标签为 1，如果不是同一个人，则标签为 0。

google 于 2015 年提出的 Facenet 中，则用到了 Triplet loss。其思想是以三元组的形式来训练模型，每次输入需要三张图片，其中两张图片是同一个人，以及一张其他人的图片，要求同一个人的照片之间的距离要小于不同人之间的距离，且要超过一个 margin。具体公式如下所示：

$$L_{triplet} = \sum_{i=1}^{N} max\{0, \| f(x_i^a) - f(x_i^p) \|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha\}$$

其中，$f(\cdot)$ 表示输入图像的特征表示，$x_i^a$ 表示 anchor 样本，$x_i^p$ 表示 positive 样本，$x_i^n$ 表示 negative 样本，$\alpha$ 表示对应的 margin。

基于 margin 的分类方法 (margin based classification)

第二类思想是基于分类的思想来进行特征提取，根据训练集中的数据，同一个人的照片属于同一类，训练集一共包含多少个 id，则总共需要分多少类。由于用的分类的思想，所以自然而然可以使用分类的损失函数。而在此基础上，又提出了一系列的方法，用以最小化类内间距或者最大化类间间距。比较有代表性的方法有 Center loss[109]，SphereFace[110]，CosFace[111] 以及 ArcFace[112] 等。

由于是多分类任务，所以最基本的损失函数形式是 softmax loss，softmax loss 是指概率输出函数为 softmax 函数，且损失函数为交叉熵 (cross entropy)。但是直接用 softmax loss 训练出来的特征，往往效果不理想，某些类别的类内间距甚至比类间间距大，导致人脸识别的时候出现错误。center loss 引入了类内中心，为每个类别提供一个类内中心，最小化训练集中每个样本与其类内中心的距离，从而达到减少类内间距的效果。具体形式如下所示：

$$\begin{aligned} L_{centerloss} &= L_S + \lambda L_C \\ &= -\frac{1}{N}(\sum_{i=1}^{N} log(\frac{e^{W_{y_i}^T f(x_i) + b_{y_i}}}{\sum_{j=1} e^{W_j^T f(x_i) + b_j}}) + \frac{\lambda}{2} \sum_{i=1}^{N} \| f(x_i) - c_{y_i} \|_2^2) \end{aligned}$$

其中，$c_{y_i}$ 表示 $x_i$ 所对应类别的中心，它与 $f_(x_i)$ 的维度一致，该类别中心与参数一样，需要在训练中迭代更新。

基于 SphereFace 的训练方式，是在此基础上做了改进，对权重进行了归一化，且增加了角度裕量，在 cos 函数上对角度乘上因子 m，加大分类难度。具体形式如下所示：

$$L_{sphereface} = -\frac{1}{N} \sum_{i=1}^{N} log(\frac{e^{\|f(x_i)\|_2 cos(m\theta_{y_i,i})}}{e^{\|f(x_i)\|_2 cos(m\theta_{y_i,i})} + \sum_{j \neq y_i} e^{\|f(x_i)\|_2 cos(\theta_{j,i})}})$$

CosFace 和 ArcFace 更进一步，对特征 $f(x_i)$ 也做了归一化，并分别引入了不同的 margin 形式，取得了更好的效果，如下所示：

$$L_{cosface} = -\frac{1}{N}\sum_{i=1}^{N} log(\frac{e^{s(cos(\theta_{y_i,i})-m)}}{e^{s(cos(\theta_{y_i,i})-m)} + \sum_{j\neq y_i} e^{scos(\theta_{j,i})}})$$

$$L_{arcface} = -\frac{1}{N}\sum_{i=1}^{N} log(\frac{e^{scos(\theta_{y_i,i}+m)}}{e^{scos(\theta_{y_i,i}+m)} + \sum_{j\neq y_i} e^{scos(\theta_{j,i})}})$$

以上方法都是通过不同方式去减少同一个人的类内间距以及增大不同人之间的类间间距。但除了损失函数以及网络的设计之外，更为重要的是训练的数据的分布，比较好的训练数据是同一个人包含多张不同的照片，这些照片覆盖此人不同年龄段，不同姿态角度，不同遮挡程度，不同妆容情况等，这样的数据能够学习到鲁棒性更强，通用性更好的模型。

## 5.3 应用案例

人脸识别目前在安防、金融等领域都得到了广泛应用，下面介绍一些常见的应用案例。

### 门禁闸机

这是人脸识别的典型应用场景，属于人脸检索（1：N）的应用。门禁闸机在初始化的时候，会要求录入一个人脸库，该人脸库经过特征提取后，作为识别的底库。当有人通过闸机的时候，会拍摄来人的照片，通过特征提取转化为特征表示之后，与底库中的特征集合进行对比，找出该人员是否存在于底库中。

### 金融核身业务

目前几乎所有的金融核身业务都支持人脸核身功能，属于人脸比对（1:1）的应用。当用户的办理某些业务的时候，会被要求进行人脸核身，系统会通过摄像头采集用户的照片，与用户留底的另一张照片进行比对，以确定用户是否为本人。这种方式大大减少了金融业务中进行业务审核的人员数量及审核时间，节省了用户时间，提升了用户体验。

监控系统

在银行、机场、商场等公共场景对人群进行监控，这种场景属于非配合式场景，用户大多数情况下并不知道有摄像头在拍摄，不会对拍摄做出配合。因此此类场景的拍摄距离通常会较远，用户姿态多样，且可能会有各种遮挡问题。相比于上面两种应用场景，此类场景较复杂，技术难度也较高。这种场景下，除了对监控中的人进行人脸检索外，还可能会进行人脸属性识别，如识别人的年龄，性别等，以便于后续进行更细致的分析。

# Chapter 6
# 声纹识别

本章简单介绍声纹识别的概念，并其常见的算法，以及在实际中的应用。

## 6.1 问题定义

### *6.1.1* 基本定义

声纹识别（Voice Print Recognition），也称作说话人识别（Speaker Recognition），是一种生物识别技术，能够根据说话人的声音特征提供精准、高效、便捷的身份识别服务。从感官直觉上来说，声纹虽然不像人脸、指纹的个体差异那样直观可见，但由于人在讲话时使用的发声器官–舌、牙齿、喉头、肺、鼻腔在尺寸和形态方面每个人的差异很大，因此反映到任何两个人的声纹图谱都存在有差异。最直观的感受是当我们打电话给认识的人的时候，通过很短一句话甚至一声“喂？”，就能准确地分辨出接电话的是谁。这种语音中承载的说话人身份信息的唯一性使得声纹也可以像人脸、指纹那样作为生物信息识别技术的生力军，可广泛应用于金融安全，公共安防，智能家居等领域。

### *6.1.2* 分类

声纹识别通常分为两大类，说话人确认和说话人辨别，也就是常说的声纹 1：1 识别和声纹 1：N 识别。声纹 1：1 识别是指确认某段语音是否是指定的某个人所说的，而声纹 1：N 识别是判断某段语音是若干人中的哪一个所说的。不同的任务和应用会使用不同的声纹识别技术，如诈骗电话需要缩小人员范围时可能需要声纹 1：N 技术进行辨别，而银行金融交易时则需要声纹 1：1 识别技术进行确认。

从另一方面，声纹识别也分为文本相关（Text-Dependent）和文本无关（Text-Independent）两类。与文本相关的声纹识别要求用户按规定内容进行朗读，这样能更加准确的建立模型，在识别的时候也要求用户按规定内容朗读。文本无关的声纹识别并不要求用户根据指定文本进行朗读，这样也能建立模型，验证的时候同样不需要用户根据指定文本进行朗读。一般来说，文本相关的声纹识别效果会更好，安全性更高，但是用户体验较差和使用场景就相对较窄，通常用于安全要求比较高的场景，如金融核实身份。文本无关的声纹识别会和文本依赖比较弱，因此能进行跨语种使用，就算我们没有别的语种的语料，也是能应用到那语种上面去。文本无关的使用场景很宽广，在金融客服上就能建立声纹库，后面的用户在金融上快速验证和关联。

### *6.1.3* 挑战和机遇

声纹识别技术的应用会有很多挑战，比如同一个人的声音具有易变性，可以控制不同部分造成不同发音；同时会受到身体状况、年龄、情绪等的影响，对识别性能有影响；又比如不同的麦克风、信道、环境噪音等对识别性能有比较大的干扰；又比如混合说话人的情形下人的声纹通常不易分离和识别等等。当然这些挑战也会被利用转化为优势，比如声音会受身体状况，年龄，情绪的影响，因此会单独训练出利用声纹识别情绪，或者利用声纹识别是否生病目前有应用到养猪场识别病猪等等。

尽管如此，与其他生物特征相比，声纹识别的应用有一些特殊的优势：

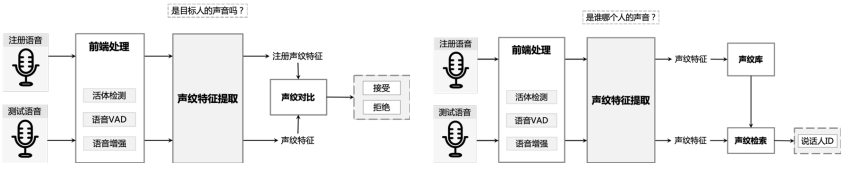
- 获取声纹数据成本较低

- 用户使用接受程度较高
- 使用成本较低，适合远程身份确认
- 随着文本无关技术提升，能在小语种、跨种族等方向都有应用场景

这些优势受到系统开发者和用户青睐，从而使得声纹识别的应用越来越受到欢迎。

## 6.2 实现方案

声纹识别中通常的流程如图 6-1所示。不管是 1：1 还是 1：N 大致都是分为三部分：前置处理，特征抽取和声纹匹配。前置处理通常是 VAD 检测、反欺诈活体检测、声音增强等。VAD 检测处理是把有声音部分和静音部分区分出来，把有声音部分送到后续处理中去。反欺诈活体检测是主要应对声纹识别是否被攻击，提高安全性。这些前置处理不在这里详细展开细说，匹配算法我们可以用最普通的计算向量距离如欧式距离或 cos 值，也可以用深度学习网络进行计算相似度。接下来我们重点介绍声纹的特征抽取。




(a) 声纹 1：1 识别　　(b) 声纹 1：N 识别

**图 6-1** 声纹识别流程

### *6.2.1* 特征抽取

声纹识别的特征抽取大致经历了三代算法，GMM 模型到 ivector 最后到深度学习网络。我们重点介绍前面两代，深度学习网络的方法和人脸特征抽取方法类似，也是分为两大类：度量学习 (Metric Learning) 和

基于 margin 的分类方法 (Margin Based Classification)，将会在人脸识别中重点介绍。

#### 6.2.1.1 GMM-UBM 模型

GMM 模型 [38] 即高斯混合模型，是由多个高斯函数进行加权求和进行拟合复杂的函数，如

$$p(x) = \sum_{k=1}^{K} p(k)p(x|k) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{6.1}$$

其中，$\pi_k$ 表示第 $k$ 个高斯函数的权重，$\mu_k$ 和 $\Sigma_k$ 表示第 $k$ 个高斯的均值和方差。

GMM 模型的参数求解，一般使用 EM 算法进行求解。通常情况下，GMM 模型可以平滑地逼近任意形状的函数，具备对实际数据极强的表征力。而声纹识别实际上就是从不同语音中抽取出相同的表征特征来。GMM 模型同时还具备比较好的泛化能力。因此 GMM 模型在声纹识别初期获得比较好的效果。随着 $k$ 的增大，所需要的训练数据也就更加大了，否则获取不到泛化能力较好的模型。

在实际使用过程中，每个人语音数据有限，很难获取到比较通用的声纹识别模型。为了解决这个问题，DA Reynolds 提出了通用背景模型 (Universal Background Model) [39]，简称 UBM。先使用大量和说话人无关的语音数据训练一个 GMM 模型，然后再使用少量的说话人数据，通过自适应算法 (如最大后验概率 MAP、最大似然线性回归 MLLR 等) 获取到说话人的个性特征的模型叫做 UBM 模型。这个思想有点像现在深度学习的 finetune 思想。这个模型就是 GMM-UBM 模型。该模型参数可以减半并有更快收敛的特点。

随着实际应用，GMM-UBM 的存在问题：参数仍然很大和受信道的干扰比较大。学术界提出了 GMM-SVM 模型 [40]、JFA 模型 [41] 等等去优化解决。

#### 6.2.1.2 ivector 模型

基于 GMM-UBM 的模型，基本是基于特征声纹空间与特征信道空间的独立假设，但是在现实使用中，数据之间都是具有相关性的。之前的假设更多是方便了公式推导同时也限制了模型的泛化能力。N.Dehak 认为既然声纹信息与信道信息不能做到完全独立，那就用一个一段低维度的定长向量同时描述声纹信息和信道信息，从而提出了 ivector 模型 [42]。

对于每一段语音都有高斯均值向量 $M$ 表示如下：

$$M = m + \omega T \tag{6.2}$$

其中，$m$ 表示通用背景模型 (UBM) 的高斯均值向量，该值和声纹信息、信道信息无关，$\omega$ 是全局差异空间因子，即为 ivector 向量，它的先验服从标准正态分布 $N(0,1)$，$T$ 表示全局的差异空间矩阵。接下来只需要估计 $\omega$ 和 $T$ 值即可。

对于 $\omega$ 和 $T$ 的参数估计，我们基于假设：每一段语音都来自不同的说话人。首先计算训练数据中每个说话人所对应的 Baum-Welch 统计量，随后随机产生 T 的初始值。后续采取 EM 算法估算得到相关的参数。

## 6.3 应用案例

### *6.3.1 声纹 1：1 识别应用案例*

#### 6.3.1.1 电话客服核身

在电话客服中应用声纹核身，可以节省核身时间，降低运营成本；并减少核身问题，提升客户体验；更重要的是，即使犯罪分子掌握了用户的所有信息，也能通过声音判断是否为本人，是否存在欺诈风险。例如，部分银行的服务热线目前已接入声纹核身技术，大致流程如图 6-2。

#### 6.3.1.2 社保核查

声纹验证能够解决参保人员面临的远程和现场身份核查及生存验证的问题，避免了指纹验证和人脸识别等需要现场办理、不易采集、伪造等

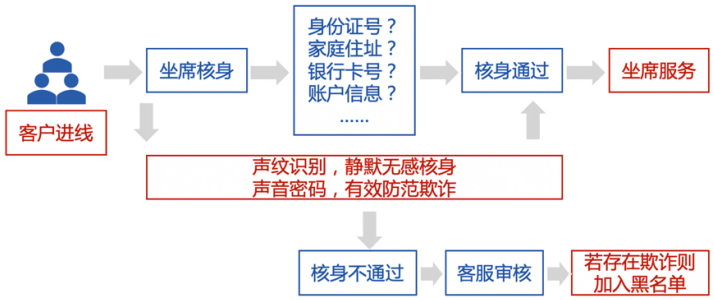



**图 6-2** 电话客服核身基本流程

问题，有效杜绝冒领养老金的可能性，节约社保资金和人力成本。例如，印尼新一代养老金认证系统已通过声纹验证技术，使其 250 万离退休人员在领取养老金时可通过电话或手机 app 进行远程身份认证，不仅节省了大量人力投入，还显著降低了传统骗保率。

#### 6.3.1.3 声纹锁

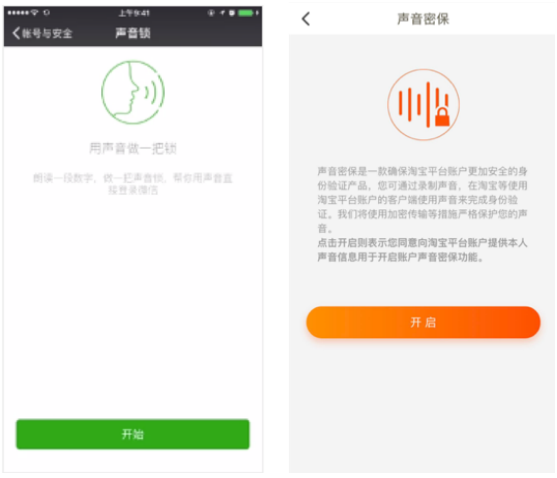
如图 6-3所示声纹验证可作为登录、密保、修改账户信息的一种验证方式，也可以应用于门禁/闸机等。

#### 6.3.1.4 声纹支付

央行发布声纹识别安全应用技术标准，认定声纹技术可适用于手机银行、第三方支付，通过与语音交互硬件结合，能够解决无屏或远场进行身份验证的痛点。

#### 6.3.1.5 声纹唤醒

通过声纹识别“主人”身份，只允许主人唤醒设备，或定制个性化服务，可适用于手机助手、智能音箱、智能家居、车载助手、服务机器人等智能设备。如图 6-5所示，在车联网应用中用户可以提前注册声纹信息并添加




(a) 声纹锁（一） (b) 声纹锁（二）

**图 6-3** 声纹锁样例


(a) 声纹支付 (b) 声纹支付界面

**图 6-4** 声纹支付样例

个性化配置，车机将通过声纹识别确认当前的驾驶人身份，可快速切换至对应的用户配置，令行车体验更加轻松。


**图 6-5** 汽车声纹唤醒

### *6.3.2 声纹 1: N 识别应用案例*

#### 6.3.2.1 公安

以破案、追逃为导向，利用声纹识别技术公安可进行“案查人”、“人查案”、“案查案”与“人查人”等多种排查方式：

- 案查人：如电信诈骗，主要线索只有语音的情况下，将该语音进行声纹库大库检索，快速锁定嫌疑人。
- 人查案：公安抓捕到可疑人员后，提取出该人的声纹特征，将其放入尚未侦破的语音案件中，排查该人是否为在逃人员。
- 案查案：公安人员可使用声纹识别技术将尚未侦破的语音案件以及语音线索归纳整理，从中排查是否有多起案件是同一人所为，帮助侦察人员获得更多线索，提高排查效率。
- 人查人：公安机关在抓捕到可疑人员后，提取出该人的声纹特征，为避免该人使用伪造身份，可将其声纹特征放入已知人员的声纹库，查询其真实身份。

声纹识别技术还能应用于重点人员监管、反电信诈骗、反恐、刑事案件侦破、身份查询与核验，助力公安有效遏制与打击犯罪，构建和强化安全的社会公众环境。例如：

- 反电信网络欺诈：在通信系统或安全监测系统中嵌入声纹识别技术，能够对黑名单人员语音对话实时预警，提示重点人员可疑行为；语音内容关键词识别动态预警，提示可疑案件与犯罪意图。

- 动态声纹布控：通过声纹识别和声纹大数据技术，进行对重点人员和关键卡口的布控监管，在第一时间完成举报人或嫌疑人身份鉴定，辅助刑事案件侦破和案情分析。


**图 6-6** 公安布控监管系统

#### 6.3.2.2 金融黑名单识别

将信贷黑名单、风险等级高、不良中介、金融欺诈等用户声纹加入黑名单库，当其再次办理业务时，匹配到黑名单库的用户，直接给出风险预警。例如，车险业务能够针对报假案、修理厂、黑中介等不良用户建立黑名单声纹库，当不良用户再次报案时，业务员端能够及时给出预警。

#### 6.3.2.3 客户定制化服务

对于某些小群体用户，如银行 VIP 客户电话呼入时，可通过声纹 1:N 匹配 VIP 声纹库，识别用户身份，从而进行定制化服务。

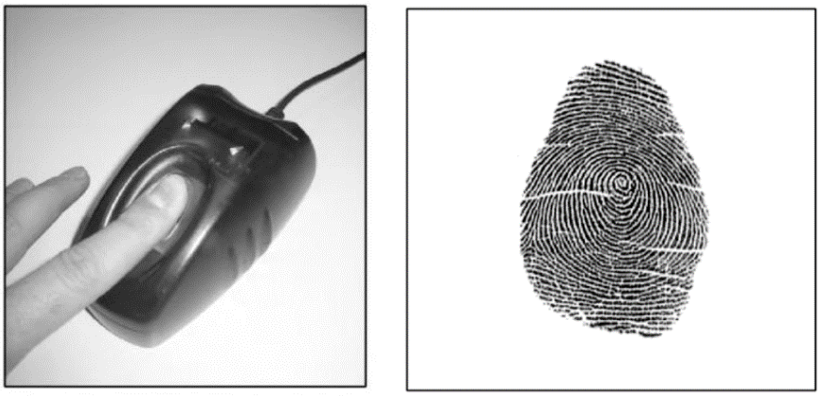
# Chapter 7
# 其他生物特征识别

随着移动互联网、智能移动终端设备的快速发展，由于具备便捷的使用体验、可靠的安全保障，以人脸识别为代表的生物特征识别技术 (biometrics) 得到了迅速应用和推广。生物特征识别技术是利用固有的生物特征进行身份认证的一类技术，由于生物特征通常具备唯一性，如果具备可测量和可验证，那么利用生物识别技术进行身份认证往往安全、可靠和准确。除了前面介绍的人脸识别之外，下面将介绍几种在金融行业应用的生物识别技术以及应用场景。

## 7.1 指纹/掌纹识别

作为生物识别技术在金融领域应用最早的一项技术，指纹识别早在上世纪 90 年代就大规模进军金融行业。指纹是人类手指末端由凹凸的皮肤所形成的纹路，如图7-1(b)所示，每个个体指纹的形状不会随着年龄发生改变，而且每个人的指纹都是不同的。指纹识别技术通过分析所采集的指纹图像中可测量的特征点并提取特征值，然后进行比对认证。指纹识别目前也早已在消费电子、安防等领域广泛应用，相关针对指纹的国家标准也已陆续制定和发布，指纹识别的性能已得到明显提高，技术也最为成熟，目前已处于应用成熟的平台期。

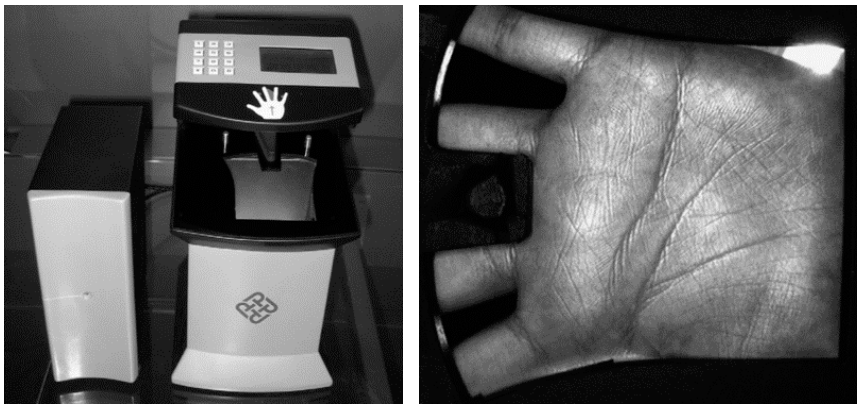
掌纹识别是近些年提出的一种相对较新的生物特征识别技术。掌纹一般指手指末端到手腕之间这一区域的手掌表面的各种纹理特征，如图7-2(b)所示。与指纹识别类似，每个人的掌纹纹理都不一样，掌纹中具有的



(a) 指纹采集设备 (b) 指纹图像

**图 7-1** 指纹采集设备和图像 [113]

很多特征可以进行测量并提取特征值，进而进行身份认证。指纹识别和掌纹识别都是非侵犯性的识别方法，实际应用中用户接受度较高。

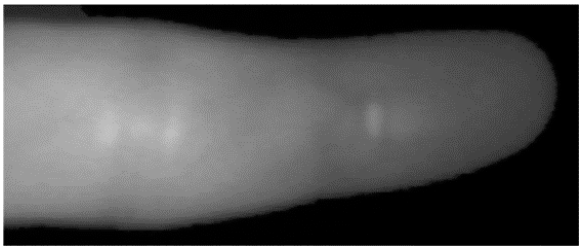


(a) 掌纹采集设备 (b) 掌纹图像

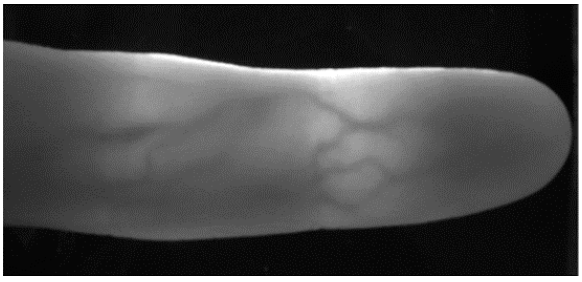
**图 7-2** 掌纹采集设备和图像 [114]

## 7.2 静脉识别

静脉识别利用的是静脉中的血红蛋白相对于肌肉、骨骼等其他生理组织对近红外光的吸收率更高，当有近红外光照射在手指或者手掌上时，通过红外摄像头获取手指或手掌的图像，静脉血管会呈现深色，肌肉组织则为浅色，呈现出黑白对比分明的图像特点，如图7-3所示，静脉血管结构可以清晰的得到呈现。静脉识别技术通过算法对图像进行分析，提取特征值进行身份认证。静脉采集设备按照近红外光源和图像传感器的相对位置不同，主要分为透射式和反射式两种，如图7-4所示。



(a) 指静脉反射式成像



(b) 指静脉透射式成像

**图 7-3** 指静脉图像 [115]

由于静脉属于人体内部特征，相比人脸和指纹来说极难复制和盗取，且只有在活体上才能采集到，因此这项技术的安全性更好，更难以被盗取和假冒。静脉识别过程中一般受外界环境因素（例如温度、湿度等）以及个体皮肤表面状态（如粗糙程度、是否磨损等）影响较小，可靠性较高。同时在使用过程中，用户手无需与设备进行接触，这种非接触式的使用更加卫生、易于用户接受。目前静脉识别技术发展时间较短，识别准确率

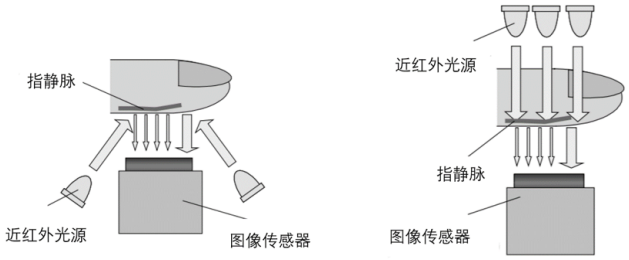



**图 7-4** 两种指静脉图像采集方式示意图 [115]

有望进一步得到提升。由于在活体鉴定方面的优势，与指纹融合可以较好地预防假指纹攻击，并提高识别准确率。静脉识别在金融行业的应用也已开始了探索。2016 年 1 月，工农中建四大银行委托广电运通起草指静脉在金融行业的应用标准。2016 年 11 月，广东省社保基金管理局也已经制定指静脉在社保行业的应用标准，计划全省开始推广指静脉养老金发放的认证工作。

## 7.3 虹膜识别

虹膜是位于人眼表面黑色瞳孔和白色巩膜之间的圆环状区域，在近红外光下能够呈现出丰富的纹理，如图7-5所示。而且虹膜在胎儿发育阶段形成后的整个生命历程中是保持不变的。这些特征决定的虹膜特征以及用于身份识别的唯一性。虹膜识别属于非接触式识别，通过专门的虹膜图像采集装置采集清晰的虹膜图像提取特征进行身份认证，识别过程高效且准确率高。虹膜识别技术被认为是生物特征识别技术中准确率最高的技术之一，在金融领域一般应用于金库管理、押运管理的较多，通过虹膜识别确认出入和押运人员身份，确保财产安全；同时，也有部分银行在尝试将虹膜识别和指静脉识别集成于自助终端中，实现更高安全级别的身份认证，以帮助用户完成自助贷款、自助理财等业务的办理。在其他行业，虹膜识别技术因眼镜、光线干扰和特征部位与采集方式等因素，目前主要用于煤矿工人等其他种类生物特征难以采集和识别的人群。

以上是目前较为常见的几种生物特征识别技术。随着智能终端设备以及生物特征传感器的快速普及和优化，生物特征识别技术已经进入大

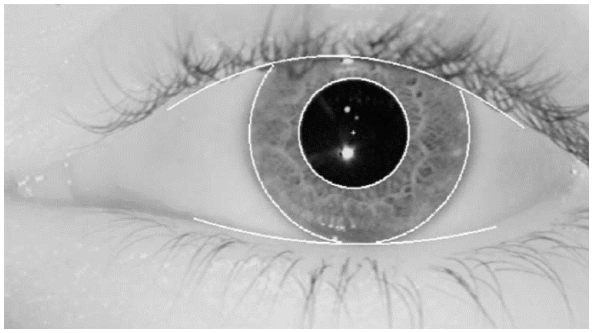


**图 7-5** 虹膜图像 [116]

规模应用阶段。单一的人脸识别或指纹识别已经难以满足金融机构的多样化需求，多模态生物识别是金融科技不可更改的趋势。

# Chapter 8
# 反欺诈：声纹与人脸识别的抗攻击

这是里关于主题模型和语言模型的介绍。

## 8.1 声纹识别中的抗攻击

声纹识别中的抗攻击

## 8.2 人脸识别中的抗攻击

人脸活体检测（Face Anti-Spoofing）技术是人脸识别系统中，用以确认待认证对象是否为真实生物活体的一项技术。一方面，人脸识别技术的商业化愈加成熟和广泛，极大改善和推进了社会金融活动的智能化和便捷性；另一方面，由于人脸照片、视频数据相对容易获取和复制，若无活体检测这一环节，那么使用被盗取的合法用户的照片、视频或者 3D 面及头套等即可入侵人脸识别系统，由此将带来极大危害 [117]。在目前的人脸识别系统中，常见的活体检测技术包括动作活体、3D 活体、红外活体、光线活体等。下面将逐一简单介绍。

动作活体是通过利用人脸关键点和人脸跟踪等技术，检测用户眨眼、张嘴、摇头等多种动作及其组合，可有效抵御照片、换脸、面具、遮挡以及屏幕翻拍等常见的攻击手段，但较难抵御视频回放攻击。

3D 活体通过专用硬件（例如 3D 结构光、ToF 等）获取人脸部 3D 结构信息，可以有效防御如手机、电脑等屏幕显示和打印照片等 2D 攻击手段，但是需要配合其他方法抵御近几年出现的高质量 3D 面具攻击 [118]。

红外活体检测一般利用人体皮肤对近红外光的反射率较高，相比于其他材质有明显区别的特性，通过专用红外设备获取人脸部红外图像判断是否为活体。实际应用中通常使用主动红外摄像，即通过红外 LED 照射人脸，利用红外摄像头获取人脸部图像，分类判断是否为活体。红外活体检测技术对于常见攻击手段具有较好的防御效果，缺点在于需要特定红外设备 [118, 119]。

光线活体是近两年出现的一种活体检测技术。由于 3D 活体、红外活体需要特殊设备，在已有系统中部署较为困难。光线活体技术利用屏幕发出不同颜色和强度的光线照射在人脸上，由于人脸自身的三维结构以及皮肤等生理组织对于不同颜色光线的反射率不同，从获取的视频中提取相应的活体信息，如图8-1所示。这项技术由于无需特殊硬件设备、且具有较高的准确率，在手机等移动端使用较为方便。其缺点在于要求视频拍摄过程稳定，闪光带来的用户体验需要得到提升，同时户外强光也会带来较大干扰 [120]。

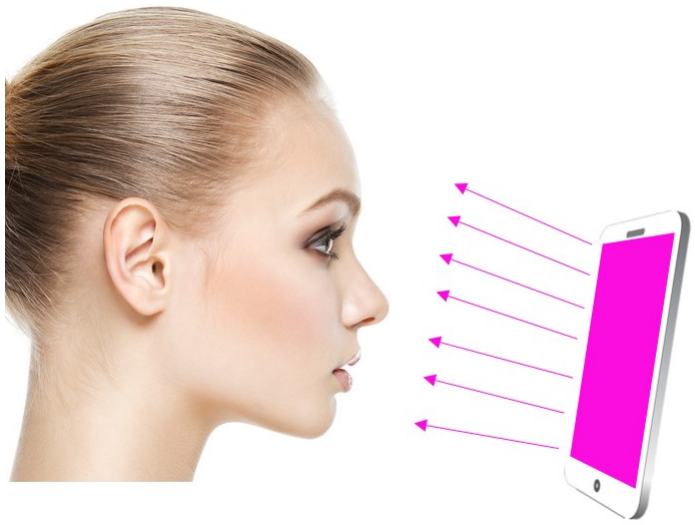


**图 8-1** 光线活体

## References

[1] MARKOV A A. An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains. Classical text in translation[J]. Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg, 1913, 23 : 591 – 600.

[2] SHANNON C E. A mathematical theory of communication[J]. Bell system technical journal, 1948, 27(3) : 379 – 423.

[3] CHOMSKY N. Three models for the description of language[J]. IRE Transactions on information theory, 1956, 2(3) : 113 – 124.

[4] STRATONOVICH R L. Conditional markov processes[G] // Nonlinear transformations of stochastic processes. [S.l.] : Elsevier, 1965 : 427 – 453.

[5] JAYNES E T. Information theory and statistical mechanics[J]. Physical review, 1957, 106(4) : 620.

[6] VITERBI A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm[J]. IEEE transactions on Information Theory, 1967, 13(2) : 260 – 269.

[7] KUČERA H, FRANCIS W N. Computational analysis of present-day American English[M]. [S.l.] : Dartmouth Publishing Group, 1967.

[8] BAKER C F, FILLMORE C J, LOWE J B. The berkeley framenet project[C] // Proceedings of the 17th international conference on Computational linguistics-Volume 1. 1998 : 86 – 90.

[9] LAFFERTY J, MCCALLUM A, PEREIRA F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J], 2001.

[10] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C] // Proceedings of the 40th annual meeting on association for computational linguistics. 2002 : 311 – 318.

[11] COLLINS M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms[C] // Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. 2002 : 1 – 8.

[12] PANG B, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques[C] // Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. 2002 : 79 – 86.

[13] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan) : 993 – 1022.

[14] TASKAR B, KLEIN D, COLLINS M, et al. Max-margin parsing[C] // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004 : 1 – 8.

[15] HOVY E, MARCUS M, PALMER M, et al. OntoNotes: the 90% solution[C] // Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers. 2006 : 57 – 60.

[16] WITTEN I H, MILNE D N. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links[J], 2008.

[17] GO A, BHAYANI R, HUANG L. Twitter sentiment classification using distant supervision[J]. CS224N project report, Stanford, 2009, 1(12) : 2009.

[18] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb) : 1137 – 1155.

[19] COLLOBERT R, WESTON J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C] // Proceedings of the 25th international conference on Machine learning. 2008 : 160 – 167.

[20] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[21] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.

[22] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404.2188, 2014.

[23] GRAVES A, JAITLY N, MOHAMED A-R. Hybrid speech recognition with deep bidirectional LSTM[C] // 2013 IEEE workshop on automatic speech recognition and understanding. 2013 : 273 – 278.

[24] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C] // Proceedings of the 2013 conference on empirical methods in natural language processing. 2013 : 1631 – 1642.

[25] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C] // Advances in neural information processing systems. 2014 : 3104 – 3112.

[26] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[27] GRAVES A, WAYNE G, DANIHELKA I. Neural turing machines[J]. arXiv preprint arXiv:1410.5401, 2014.

[28] WESTON J, CHOPRA S, BORDES A. Memory networks[J]. arXiv preprint arXiv:1410.3916, 2014.

[29] SUKHBAATAR S, WESTON J, FERGUS R, et al. End-to-end memory networks[C] // Advances in neural information processing systems. 2015 : 2440 – 2448.

[30] KUMAR A, IRSOY O, ONDRUSKA P, et al. Ask me anything: Dynamic memory networks for natural language processing[C] // International conference on machine learning. 2016 : 1378 – 1387.

[31] GRAVES A, WAYNE G, REYNOLDS M, et al. Hybrid computing using a neural network with dynamic external memory[J]. Nature, 2016, 538(7626) : 471 – 476.

[32] HENAFF M, WESTON J, SZLAM A, et al. Tracking the world state with recurrent entity networks[J]. arXiv preprint

arXiv:1612.03969, 2016.

[33] CONNEAU A, KIELA D, SCHWENK H, et al. Supervised learning of universal sentence representations from natural language inference data[J]. arXiv preprint arXiv:1705.02364, 2017.

[34] MCCANN B, BRADBURY J, XIONG C, et al. Learned in translation: Contextualized word vectors[C] // Advances in Neural Information Processing Systems. 2017 : 6294 – 6305.

[35] SUBRAMANIAN S, TRISCHLER A, BENGIO Y, et al. Learning general purpose distributed sentence representations via large scale multi-task learning[J]. arXiv preprint arXiv:1804.00079, 2018.

[36] RAMACHANDRAN P, LIU P J, LE Q V. Unsupervised pretraining for sequence to sequence learning[J]. arXiv preprint arXiv:1611.02683, 2016.

[37] HOWARD J, RUDER S. Universal language model fine-tuning for text classification[J]. arXiv preprint arXiv:1801.06146, 2018.

[38] REYNOLDS D A. Speaker identification and verification using Gaussian mixture speaker models[J]. Speech communication, 1995, 17(1-2) : 91 – 108.

[39] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital signal processing, 2000, 10(1-3) : 19 – 41.

[40] CAMPBELL W M, STURIM D E, REYNOLDS D A, et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation[C] // 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings : Vol 1. 2006 : I – I.

[41] KENNY P. Joint factor analysis of speaker and session variability: Theory and algorithms[J]. CRIM, Montreal,(Report) CRIM-06/08-13, 2005, 14 : 28 – 29.

[42] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4) : 788 – 798.

[43] WONG P-K, CHAN C. Chinese word segmentation based on maximum matching and word binding force[C] // Proceedings of the 16th conference on Computational linguistics-Volume 1. 1996 : 200 – 203.

[44] WEICHUN H, JIANJIAN J. Research on Longest Backward Segmentation for Context[C] // 1st International Workshop on Cloud Computing and Information Security. 2013.

[45] YANG L, XU L, SHI Z. An enhanced dynamic hash TRIE algorithm for lexicon search[J]. Enterprise Information Systems, 2012, 6(4) : 419 – 432.

[46] LI Q-H, CHEN Y-J, SUN J-G. A New Dictionary Mechanism for Chinese Word Segmentation [J][J]. Journal of Chinese Information Processing, 2003, 4 : 001.

[47] NG H I, LUA K T. A word finding automation for Chinese sentence tokenization[J]. submitted to ACM Transaction of Asian Languages Processing, .

[48] ZHANG H-P, LIU Q, CHENG X-Q, et al. Chinese lexical analysis using hierarchical hidden markov model[C] // Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. 2003 : 63 – 70.

[49] KUPIEC J. Robust part-of-speech tagging using a hidden Markov model[J]. Computer speech & language, 1992, 6(3) : 225 – 242.

[50] YU H-K, ZHANG H-P, LIU Q, et al. Chinese named entity identification using cascaded hidden Markov model[J]. Journal-China Institute Of Communications, 2006, 27(2) : 87.

[51] MORWAL S, JAHAN N, CHOPRA D. Named entity recognition using hidden Markov model (HMM)[J]. International Journal on Natural Language Computing (IJNLC), 2012, 1(4) : 15 – 23.

[52] MCCALLUM A, FREITAG D, PEREIRA F C. Maximum Entropy Markov Models for Information Extraction and Segmentation.[C] // Icml : Vol 17. 2000 : 591 – 598.

[53] RATNAPARKHI A. A maximum entropy model for part-of-speech tagging[C] // Conference on Empirical Methods in Natural Lan-

guage Processing. 1996.

[54] BORTHWICK A, GRISHMAN R. A maximum entropy approach to named entity recognition[D]. [S.l.] : Citeseer, 1999.

[55] ZHAO H, HUANG C, LI M. An improved Chinese word segmentation system with conditional random field[C] // Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006 : 162 – 165.

[56] CONSTANT M, SIGOGNE A. MWU-aware part-of-speech tagging with a CRF model and lexical resources[C] // Proceedings of the workshop on multiword expressions: from parsing and generation to the real world. 2011 : 49 – 56.

[57] EKBAL A, HAQUE R, BANDYOPADHYAY S. Named entity recognition in Bengali: A conditional random field approach[C] // Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II. 2008.

[58] ZHANG Y, CLARK S. Chinese segmentation with a word-based perceptron algorithm[C] // Proceedings of the 45th annual meeting of the association of computational linguistics. 2007 : 840 – 847.

[59] ZHANG Y, CLARK S. Joint word segmentation and POS tagging using a single perceptron[C] // Proceedings of ACL-08: HLT. 2008 : 888 – 896.

[60] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(Aug) : 2493 – 2537.

[61] ZHENG X, CHEN H, XU T. Deep learning for Chinese word segmentation and POS tagging[C] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013 : 647 – 657.

[62] WU F, LIU J, WU C, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation[C] // The World Wide Web Conference. 2019 : 3342 – 3348.

[63] CHEN X, QIU X, ZHU C, et al. Long short-term memory neural networks for chinese word segmentation[C] // Proceedings of the

2015 Conference on Empirical Methods in Natural Language Processing. 2015 : 1197 – 1206.

[64] CHEN X, QIU X, ZHU C, et al. Gated recursive neural network for Chinese word segmentation[C] // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015 : 1744 – 1753.

[65] HATORI J, MATSUZAKI T, MIYAO Y, et al. Incremental joint approach to word segmentation, POS tagging, and dependency parsing in Chinese[C] // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. 2012 : 1045 – 1053.

[66] SHAO Y, HARDMEIER C, TIEDEMANN J, et al. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF[J]. arXiv preprint arXiv:1704.01314, 2017.

[67] CHANG P-C, GALLEY M, MANNING C D. Optimizing Chinese word segmentation for machine translation performance[C] // Proceedings of the third workshop on statistical machine translation. 2008 : 224 – 232.

[68] TESNIÈRE L. Eléments de syntaxe structurale[J], 1959.

[69] 黄曾阳. HNC 理论概要 [J/OL]. 中文信息学报, 1997, 11(4) : 12. http://jcip.cipsc.org.cn/CN/abstract/article_674.shtml.

[70] DOZAT T, MANNING C D. Deep biaffine attention for neural dependency parsing[J]. arXiv preprint arXiv:1611.01734, 2016.

[71] JI T, WU Y, LAN M. Graph-based dependency parsing with graph neural networks[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 : 2475 – 2485.

[72] MA X, HOVY E. Neural probabilistic model for non-projective mst parsing[J]. arXiv preprint arXiv:1701.00874, 2017.

[73] FERNÁNDEZ-GONZÁLEZ D, GÓMEZ-RODRÍGUEZ C. Left-to-right dependency parsing with pointer networks[J]. arXiv preprint arXiv:1903.08445, 2019.

[74] KIPERWASSER E, GOLDBERG Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations[J]. Transactions of the Association for Computational Linguistics, 2016, 4 : 313 – 327.

[75] MA X, HU Z, LIU J, et al. Stack-pointer networks for dependency parsing[J]. arXiv preprint arXiv:1805.01087, 2018.

[76] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods[C] // 33rd annual meeting of the association for computational linguistics. 1995 : 189 – 196.

[77] CARRERAS X, MÀRQUEZ L. Introduction to the CoNLL-2005 shared task: Semantic role labeling[C] // Proceedings of the ninth conference on computational natural language learning (CoNLL-2005). 2005 : 152 – 164.

[78] BJÖRKELUND A, HAFDELL L, NUGUES P. Multilingual semantic role labeling[C] // Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task. 2009 : 43 – 48.

[79] LUONG M-T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv preprint arXiv:1508.04025, 2015.

[80] HERMANN K M, KOCISKY T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[C] // Advances in neural information processing systems. 2015 : 1693 – 1701.

[81] SERBAN I V, SORDONI A, BENGIO Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models[C] // Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[82] CHEN H, LIU X, YIN D, et al. A survey on dialogue systems: Recent advances and new frontiers[J]. Acm Sigkdd Explorations Newsletter, 2017, 19(2) : 25 – 35.

[83] LECUN Y, BENGIO Y, OTHERS. Convolutional networks for images, speech, and time series[J]. The handbook of brain theory and neural networks, 1995, 3361(10) : 1995.

[84] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8) : 1735 – 1780.

[85] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] // Advances in neural information processing systems. 2017 : 5998 – 6008.

[86] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[J]. arXiv preprint arXiv:1607.01759, 2016.

[87] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C] // Advances in neural information processing systems. 2015 : 649 – 657.

[88] TAN J, WAN X, XIAO J. Abstractive document summarization with a graph-based attentional neural model[C] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017 : 1171 – 1181.

[89] YAO J-G, WAN X, XIAO J. Recent advances in document summarization[J]. Knowledge and Information Systems, 2017, 53(2) : 297 – 336.

[90] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C] // Advances in neural information processing systems. 2013 : 3111 – 3119.

[91] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C] // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014 : 1532 – 1543.

[92] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.

[93] SHANG L, LU Z, LI H. Neural responding machine for short-text conversation[J]. arXiv preprint arXiv:1503.02364, 2015.

[94] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv

preprint arXiv:1810.04805, 2018.

[95] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8) : 9.

[96] SONG K, TAN X, QIN T, et al. Mass: Masked sequence to sequence pre-training for language generation[J]. arXiv preprint arXiv:1905.02450, 2019.

[97] FENG M, XIANG B, GLASS M R, et al. Applying deep learning to answer selection: A study and an open task[C] // 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). 2015 : 813 – 820.

[98] ZHOU X, LI L, DONG D, et al. Multi-turn response selection for chatbots with deep attention matching network[C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 : 1118 – 1127.

[99] LU J, XIE Z, LING G, et al. Spatio-Temporal Matching Network for Multi-Turn Responses Selection in Retrieval-Based Chatbots[J], .

[100] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C] // Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 : Vol 1. 2001 : I – I.

[101] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10) : 1499 – 1503.

[102] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C] // Advances in neural information processing systems. 2015 : 91 – 99.

[103] WANG H, LI Z, JI X, et al. Face r-cnn[J]. arXiv preprint arXiv:1706.01061, 2017.

[104] TANG X, DU D K, HE Z, et al. Pyramidbox: A context-assisted single shot face detector[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018 : 797 – 813.

[105] DENG J, GUO J, ZHOU Y, et al. Retinaface: Single-stage dense face localisation in the wild[J]. arXiv preprint arXiv:1905.00641, 2019.

[106] NAJIBI M, SAMANGOUEI P, CHELLAPPA R, et al. Ssh: Single stage headless face detector[C] // Proceedings of the IEEE International Conference on Computer Vision. 2017 : 4875 – 4884.

[107] SUN Y, CHEN Y, WANG X, et al. Deep learning face representation by joint identification-verification[C] // Advances in neural information processing systems. 2014 : 1988 – 1996.

[108] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015 : 815 – 823.

[109] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition[C] // European conference on computer vision. 2016 : 499 – 515.

[110] LIU W, WEN Y, YU Z, et al. Sphereface: Deep hypersphere embedding for face recognition[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017 : 212 – 220.

[111] WANG H, WANG Y, ZHOU Z, et al. Cosface: Large margin cosine loss for deep face recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018 : 5265 – 5274.

[112] DENG J, GUO J, XUE N, et al. Arcface: Additive angular margin loss for deep face recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019 : 4690 – 4699.

[113] MALTONI D, MAIO D, JAIN A K, et al. Handbook of fingerprint recognition[M]. [S.l.] : Springer Science & Business Media, 2009.

[114] KONG A, ZHANG D, KAMEL M. A survey of palmprint recognition[J]. pattern recognition, 2009, 42(7) : 1408 – 1418.

[115] HASHIMOTO J. Finger vein authentication technology and its future[C] // 2006 Symposium on VLSI Circuits, 2006. Digest of Tech-

nical Papers.. 2006 : 5 – 8.

[116] DAUGMAN J. How iris recognition works[G] // The essential guide to image processing. [S.l.] : Elsevier, 2009 : 715 – 739.

[117] CHINGOVSKA I, ANJOS A, MARCEL S. On the effectiveness of local binary patterns in face anti-spoofing[C] // 2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG). 2012 : 1 – 7.

[118] ZHANG S, WANG X, LIU A, et al. A dataset and benchmark for large-scale multi-modal face anti-spoofing[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019 : 919 – 928.

[119] YI D, LEI Z, ZHANG Z, et al. Face anti-spoofing: Multi-spectral approach[G] // Handbook of Biometric Anti-Spoofing. [S.l.] : Springer, 2014 : 83 – 102.

[120] LIU Y, TAI Y, LI J, et al. Aurora guard: Real-time face anti-spoofing via light reflection[J]. arXiv preprint arXiv:1902.10311, 2019.