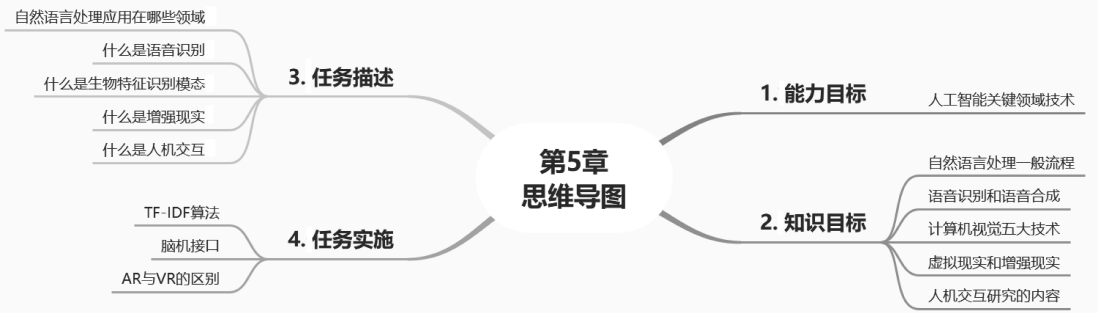
第5章

关键领域技术



5.1 自然语言处理

《圣经》里有一个故事说，巴比伦人想建造一座塔直通天堂。建塔的人都说着同一种语言，心意相通、齐心协力。上帝看到人类竟然敢做这种事情，就让他们的语言变得不一样。因为人们听不懂对方在讲什么，于是大家整天吵吵闹闹，无法继续建塔。后来人们把这座塔称为巴别塔，“巴别”的意思就是“分歧”。虽然巴别塔停建了，但一个梦想却始终萦绕在人们心中：人类什么时候才能拥有相通的语言，重建巴别塔呢？机器翻译被视为“重建巴别塔”的伟大创举。假如能够实现不同语言之间的机器翻译，我们就可以理解世界上任何人说的话，与他们进行交流和沟通，再也不必为相互不能理解而困扰。机器翻译指的是利用计算机自动地将一种自然语言翻译为另外一种自然语言。

另外，在百度或者谷歌中搜索“姚明的身高”时，搜索引擎除了给你一系列相关的网页以外，还会直接给出一个具体的答案，这就用到了自然语言问答技术。

机器翻译和问答技术都是自然语言处理领域的热点技术，在金融、教育、法律、医疗健康等领域，得到了越来越广泛的应用。

5.1.1 历史及面临的挑战

自然语言处理（Natural Language Processing，NLP）就是用计算机来处理、理解以及运

用人类语言(如中文、英文等),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,机器才算拥有了真正的智能。

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、信息检索、信息提取与过滤、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、人工智能研究以及与语言计算相关的语言学研究等。

自然语言处理兴起于美国。第二次世界大战之后,20 世纪 50 年代,当电子计算机还在襁褓之中时,利用计算机处理人类语言的想法就已经出现。1954 年 1 月 7 日,美国乔治敦大学和 IBM 公司合作实验成功地将超过 60 句俄文自动翻译成英文。虽然当时的这个机器翻译系统非常简单,仅仅包含 6 个语法规则和 250 个词,但由于媒体的广泛报道,纷纷认为这是一个巨大的进步,导致美国政府备受鼓舞,加大了对自然语言处理研究的投资。

那么,自然语言处理到底存在哪些主要困难或挑战,吸引那么多研究者几十年如一日孜孜不倦地探索解决之道呢?

一是语义理解,或者说知识的学习或常识的理解问题。这是自然语言处理技术如何变得更“深”的问题。尽管常识的理解对人类来说不是问题,但它却很难教给机器。比如我们可以对手机助手说“查找附近的餐馆”,手机就会在地图上显示附近餐馆的位置。但如果说“我饿了”,手机助手可能就无动于衷,因为它缺乏“饿了需要就餐”这样的常识,除非手机设计者把这种常识灌入到了这个系统中。但大量的这种常识都潜藏在我们意识的深处,AI 系统的设计者几乎不可能把所有这样的常识都总结出来,并灌入到 AI 系统中。

自然语言中充满了大量的歧义,人类的活动和表达十分复杂,而语言中的词汇和语法规则又是有限的,这就导致了同一种语言形式可能表达了多种不同含义。由于汉语不像英语等语言具有天然的分词,因此汉语的处理就多了分词这一层障碍。在分词过程中,计算机会在每个单词后面加入分隔符,而有些时候语义有歧义,分隔符的插入就变得困难。如“南京市长江大桥”一词,既可以理解为“位于南京的跨长江大桥”,也可以理解为“一名叫江大桥的南京市长”。要想实现正确分词,就需要结合语境,对文本语义充分理解,这显然对计算机来说是个挑战。

在短语层面上也依旧存在语言问题,例如“控制计算机”,既可以理解为动宾关系“我控制了这台计算机”,也可以理解成偏正关系“具有控制功能的计算机”。可见,如果不能正确处理各级语言单位的歧义问题,计算机就不能准确理解自然语言表达的含义。另外,上下文内容的获取问题对机器翻译来说也是一种挑战。如“我从小范手里拿走一块糖果给小李,他可高兴了。”在后一句话中,要想知道“他”指代的是小范还是小李,就要理解前一句话,小李得到糖果而小范失去了糖果,高兴的应为小李,所以“他”指代的是小李。

二是低资源问题。所谓无监督学习、Zero-shot 学习、Few-shot 学习、元学习、迁移学习等技术,本质上都是为了解决低资源问题。面对标注数据资源贫乏的问题,譬如小语种的机器翻译、特定领域对话系统、客服系统、多轮问答系统等,自然语言处理尚无良策。这类问题统称为低资源的自然语言处理问题。对这类问题,我们除了设法引入领域知识(词典、规则)

以增强数据能力之外，还可以基于主动学习的方法来增加更多的人工标注数据，以及采用无监督和半监督的方法来利用未标注数据，或者采用多任务学习的方法来使用其他任务，甚至其他语言的信息，还可以使用迁移学习的方法来利用其他的模型。这是自然语言处理技术为何变得更“广”的问题。

此外，目前也有研究人员正在关注自然语言处理方法中的社会问题，包括自然语言处理模型中的偏见和歧视，大规模计算对环境和气候带来的影响，传统工作被取代后人的失业和再就业问题等。

5.1.2 自然语言处理的一般处理流程

自然语言处理的整个流程一般可以概括为四部分，语料预处理→特征工程→模型训练→指标评价。

1. 语料预处理

(1) 数据清洗。语料是自然语言处理任务研究的内容，通常用一个文本集作为语料库。语料可以通过已有数据、公开数据集、爬虫抓取等方式获取。有了语料后，首先要做数据清洗。数据清洗，顾名思义就是在语料中找到我们感兴趣的东西，把不感兴趣的、视为噪声的内容清洗删除，包括对于原始文本提取标题、摘要、正文等信息。对于爬取的网页内容，去除广告、标签、HTML、JavaScript 等代码和注释等。常见的数据清洗方式有人工去重、对齐、删除和标注等，或者规则提取内容、正则表达式匹配、根据词性和命名实体提取、编写脚本或代码批处理等。

(2) 分词。当进行文本挖掘分析时，我们希望文本处理的最小单位粒度是词或词语，所以这个时候就需要将文本全部进行分词。常见的分词算法有基于字符串匹配的分词方法、基于理解的分词方法、基于统计的分词方法和基于规则的分词方法。每种方法下面对应许多具体的算法。

(3) 词性标注。词性标注就是给词语标上词类标签，比如名词、动词、形容词等。常用的词性标注方法有基于规则的、基于统计的算法，比如最大熵词性标注、HMM 词性标注等。词性标注是一个经典的序列标注问题，不过对于有些中文自然语言处理来说，词性标注不是非必需的。比如，常见的文本分类就不用关心词性问题，但情感分析、知识推理却是需要的。图 5.1 是常见的中文词性整理。

(4) 去停用词。停用词一般指对文本特征没有任何贡献作用的字词，比如标点符号、语气、人称等词。所以在一般性的文本处理中，分词之后，接下来一步就是去停用词。

但对于中文来说，去停用词操作不是一成不变的，停用词词典是根据具体场景来决定的，比如在情感分析中，语气词、感叹号是应该保留的，因为它们对表示语气程度、感情色彩有一定的贡献和意义。

2. 特征工程

完成语料预处理之后，接下来需要考虑的是，如何把分词之后的字和词语表示成计算机能够计算的类型。显然，如果要计算，至少需要把中文分词的字符串转换成数字，确切地说，应该是数学中的向量。有两种常用的表示模型，分别是词袋模型和词向量。

词性编码	词性名称	注　　解
Ag	形语素	形容词性语素。形容词代码为 a,语素代码 g 前面置以 A
a	形容词	取英语形容词 adjective 的第 1 个字母
ad	副形词	直接作状语的形容词。形容词代码 a 和副词代码 d 并在一起
an	名形词	具有名词功能的形容词。形容词代码 a 和名词代码 n 并在一起
b	区别词	取汉字“别”的声母
c	连词	取英语连词 conjunction 的第 1 个字母
dg	副语素	副词性语素。副词代码为 d,语素代码 g 前面置以 D
d	副词	取 adverb 的第 2 个字母,因其第 1 个字母已用于形容词
e	叹词	取英语叹词 exclamation 的第 1 个字母
f	方位词	取汉字“方”
g	语素	绝大多数语素都能作为合成词的“词根”,取汉字“根”的声母
h	前接成分	取英语 head 的第 1 个字母
i	成语	取英语成语 idiom 的第 1 个字母
i	简称略语	取汉字“简”的声母

图 5.1　常见的中文词性整理表

(1) 词袋模型。词袋模型(Bag of Word,BOW),即不考虑词语原本在句子中的顺序,直接将每一个词语或符号统一放置在一个集合(如列表),然后按照计数的方式对出现的次数进行统计。统计词频是最基本的方式,TF-IDF 是词袋模型的一个经典用法。

TF-IDF(Term Frequency-Inverse Document Frequency,词频-逆向文件频率)是一种用于信息检索(Information Retrieval)与文本挖掘(Text Mining)的常用加权技术。TF-IDF 的主要思想是:如果某个单词在一篇文章中出现的词频(Term Frequency,TF)高,并且在其他文章中很少出现,则认为此词或短语具有很好的类别区分能力,适合用来分类,如图 5.2 所示。

$$\text{词频(TF)} = \frac{\text{某个词在文章中的出现次数}}{\text{文章的总词数}} \qquad \text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}}\right)$$

图 5.2　TF-IDF

我们举一个例子,有一篇 100 字的短文,其中“猫”这个词出现了 3 次。那么这篇短文中“猫”的词频 $tf = \frac{\text{某个词在文章中出现的次数}}{\text{文章的总词数}} = \frac{3}{100} = 0.03$,如果这里有 10 000 000 篇文章,其中有“猫”这个词的文章只有 1000 篇,那么“猫”对应所有文本,也就是整个语料库的逆向文件频率 $idf = \log \frac{\text{语料库的文档总数}}{\text{包含该词的文档数}} = \frac{10\ 000\ 000}{1000} = 4$,这里 log 取 10 为底。由于 $tfidf_{i,j} = tf_{i,j} \times idf_i$,这样就可以计算得到“猫”在这篇文章中的 TF-IDF 值 $0.03 \times 4 = 0.12$。

现在假设在同一篇文章中,“是”这个词出现了 20 次,因此“是”这个字的词频为 0.2。如果只考虑词频的话,在这篇文章中明显“是”比“猫”更重要。

但我们还有逆向文件频率,假设“是”这个字在全部的 10 000 000 篇文章都出现过了,那么“是”的逆向文件频率就是 0$\left(\text{即 } \log \frac{10\ 000\ 000}{10\ 000\ 000} = 0\right)$。

这样综合来看,“是”这个字 TF-IDF 就只有 0 了,远不及“猫”重要。对于这篇文章,“猫”这个词远比出现更多次的“是”重要。诸如此类,出现很多次,但实际上并不包含文章特征信息的词还有很多,比如“这”“也”“就”“的”“了”等。

(2) 词向量。词向量是将字、词语转换成向量矩阵的计算模型。到目前为止,最常用的词向量技术是 One-hot,这种方法是把每个词表示为一个很长的向量。这个向量的维度是词表大小,其中绝大多数元素为 0,只有一个维度的值为 1,这个维度就代表了当前的词。

词向量技术还有 Google 团队的 Word2Vec,它主要包含两个模型:跳字模型(Skip-Gram)和连续词袋模型(Continuous Bag of Words,CBOW),以及两种高效训练的方法:负采样(Negative Sampling)和层序 Softmax(Hierarchical Softmax)。

以 Word2Vec 为代表的词向量技术,是自然语言处理领域一直以来最常用的文本表征方法,但这种方法仅学习了文本的浅层表征,并且这种浅层表征是上下文无关的文本表示,对于后续任务的效果提升非常有限。直到 ELMo 模型提出了一种上下文相关的文本表示方法,并在多个典型下游任务中表现惊艳,才使得预训练一个通用的文本表征模块成为可能。此后,基于 BERT 的改进模型、XLNet 等大量预训练语言模型涌出,预训练技术逐渐发展成了自然语言处理领域不可或缺的主流技术。

值得一提的是,Word2Vec 词向量可以较好地表达不同词之间的相似和类比关系。除此之外,还有一些词向量的表示方式,如 Doc2Vec、WordRank 和 FastText 等。

3. 模型训练

在选择好特征向量之后,接下来要做的事情当然就是模型训练,对于不同的应用需求,我们使用不同的模型,传统的是有监督和无监督机器学习模型,还有 KNN、SVM、Naive Bayes、决策树、GBDT、K-means 等模型,深度学习模型有 CNN、RNN、LSTM、Seq2Seq、FastText、TextCNN 等。这些模型在分类、聚类、神经网络等算法中都会讲到,这里不再赘述。

4. 指标评价

模型训练好后,在上线之前要对模型进行必要的评估,目的是让模型对语料具备较好的泛化能力。对于二分类问题,根据真实类别与学习器预测类别的组合,可把样例划分为真正例(True Positive,TP)、假正例(False Positive,FP)、真反例(True Negative,TN)、假反例(False Negative,FN)四种情形,令 TP、FP、TN、FN 分别表示其对应的样例数,显然有 TP+FP+TN+FN=样例总数。分类结果的“混淆矩阵”(Confusion Matrix)如图 5.3 所示。

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

图 5.3 分类结果的“混淆矩阵”

5.1.3 自然语言处理的主要研究方向

1. 聊天机器人

对话系统可以追溯到艾伦·图灵的图灵测试。接下来,我们学习现有聊天机器人所涉

及的技术。

（1）机器学习和深度学习。机器学习技术属于基础技术，比如说，分类算法可以用于用户的意图分类和情感分类；语言模型可以用于筛选语音识别后的句子是否通顺；聚类算法可以用于用户的行为习惯分析等。随着数据量越来越多，可以发挥深度学习的优势，更进一步提升聊天机器人的基础技术能力。

（2）自然语言处理。自然语言处理是聊天机器人语义交互层面的核心技术。比如，检索技术可以选取语料库中最合适的回复，命名实体识别可以找出句子中的关键信息，如“播放李荣浩的《李白》”中，《李白》是指一首歌名。主体识别可以用于判断句子的主语，例如“我给你唱歌”和“给我唱歌”的主语是不同的。此外，还有句型判断、实体链接、词性标注、依存分析等各项技术，综合运用于用户句子的解析。

（3）数据库技术。通过数据库技术，可以在预先存储好的大规模语料库中，快速检索相近的句子，也可以对海量的用户交互数据进行存储并进一步分析。

（4）知识图谱技术。知识图谱是聊天机器人实现认知交互的关键技术之一，可以帮助聊天机器人进行记忆、联想和推理。

（5）关于知识图谱的声学技术。关于知识图谱的声学技术包括语音识别、语音合成、声纹迁移、声纹识别以及歌声合成等，为聊天机器人提供了更加丰富的表现力。声学技术还涉及与芯片、硬件（例如麦克风阵列）的配合。

（6）计算机视觉技术。通过计算机视觉技术，可以进行人脸识别、情绪识别，并可以进一步配合语音、语义技术对用户语句进行深度分析。

（7）其他技术。

很多聊天机器人产品具备硬件形态，包括虚拟形象，因此也需要芯片技术、硬件、全息技术、美术和设计等支持。聊天机器人一定是一个技术整合的产物，在一个有很多串行模块的系统中，有个很重要的问题是错误传递。比如说有 5 个串行模块，每个模块的性能都是 95%，而最终的结果却只有 77%。所以，在设计一个聊天机器人架构时，也需要尽可能避免模块的串行化。同时，对于多轮交互架构，也需要有更加成熟的设计。

当前，由于技术不成熟，聊天机器人还无法完全做到和人一样的聊天方式，即使市面上有很多平台可以自建聊天机器人，如微软的小冰（目前最好的闲聊）、苹果的 Siri、亚马逊的 Amazon Echo 等。

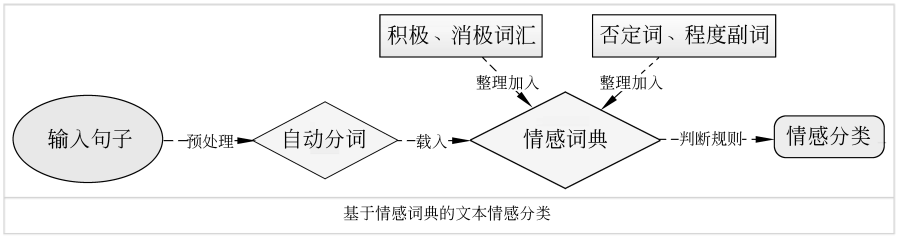
2. 情感分析

情感分析是基于自然语言处理的分类技术，主要解决的问题是判断一段话是正面的还是负面的。例如，电商类的网站根据情感分析提取正负面的评价关键词，形成商品的标签。基于这些标签，用户可以快速知道大众对这个商品的看法；还有不少基金公司会利用人们对某公司和行业的看法态度来预测未来股票的跌涨；再比如一些新闻类的网站，根据新闻的评论可以知道这个新闻的热点情况，是积极导向，还是消极导向，从而进行舆论新闻的有效控制。

情感分析可以采用基于情感词典的方法，也可以采用基于深度学习的方法。

（1）基于情感词典的方法，先对文本进行分词和停用词处理等预处理，再利用已构建好的情感词典，对文本进行字符串匹配，从而挖掘正面和负面信息，如图 5.4 所示。

情感词典在整个情感分析中至关重要，所幸现在有很多开源的情感词典，如 BosonNLP 情感词典（它是基于微博、新闻、论坛等数据来源构建的情感词典）和知网情感词典等。当然我们也可以通过语料来自己训练情感词典。



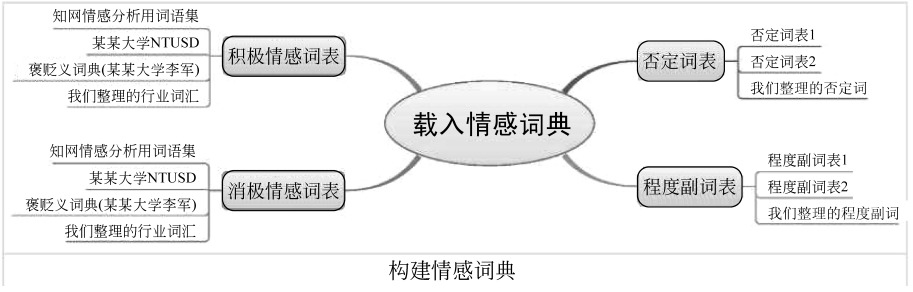


图 5.4　基于情感词典的方法

基于词典的文本匹配算法相对简单。语句分词后，逐个遍历其中的词语，如果词语命中了词典，则进行相应权重的处理。正面词权重为加法，负面词权重为减法，否定词权重取相反数，程度副词权重则和它修饰的词语权重相乘，如图 5.5 所示。

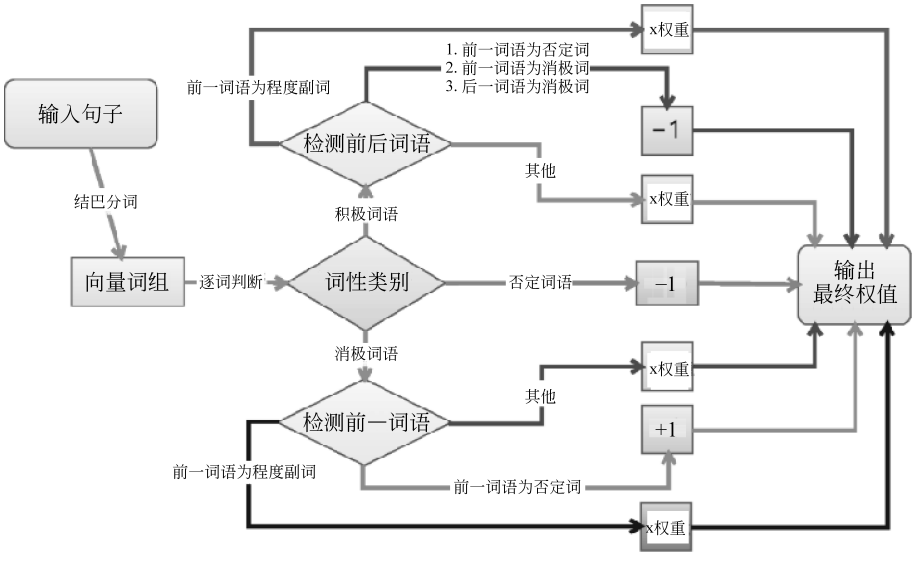


图 5.5　基于情感词典的文本分类的程序框图

利用最终输出的权重值，就可以区分是正面、负面还是中性情感了。

基于词典的情感分类，简单易行，而且通用性也能够得到保障。但仍然有精度不高、词典构建难等不足。

(2) 基于深度学习的方法。基于深度学习的方法首先对语句进行分词、停用词、简繁转换等预处理，再进行词向量编码，然后利用 LSTM 或 GRU 等 RNN 网络进行特征提取，最

后通过全连接层和 softmax 函数输出每个分类的概率，从而得到情感分类，如图 5.6 所示。

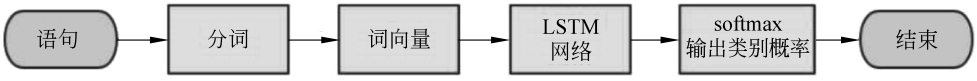


图 5.6　基于深度学习的情感分类

传统方法是人为地构造分类的特征，最终的分类效果取决于情感词库的完善性，另外还需要很好的语言学基础，也就是说，还需要知道一个句子通常在什么情况为表现为积极或消极的。深度学习方法是指选取情感词作为特征词，将文本矩阵化（转为向量），利用逻辑回归（Logistic Regression）、朴素贝叶斯（Naive Bayes）、支持向量机（SVM）等方法进行分类。最终分类效果取决于训练文本的选择以及正确的情感标注。

3. 机器翻译

机器翻译是计算机发展之初就企图解决的问题之一，目的是实现机器自动将一种语言转化为另一种语言。早期方法是语言学家手动编写翻译规则实现机器翻译，但人工设计规则的代价非常大，对语言学家的翻译功底要求非常高，并且规则很难覆盖所有的语言现象。之后 IBM 公司在 20 世纪 90 年代提出了统计机器翻译的方法，这种方法只需要人工设计基于词、短语和句子的各种特征，提供足够多的双语语料，就能相对快速地构建一套统计机器翻译系统（Statistical Machine Translation，SMT），大大减少了翻译系统设计研发的难度，翻译性能也超越了基于规则的方法。于是，机器翻译也从语言学家主导转向计算机科学家主导，在学术界和产业界中基于统计的方法也逐渐取代了基于规则的方法。随着深度学习不断在图像和语音领域的各类任务中达到最先进水平，机器翻译的研究者也开始使用深度学习技术。

自然语言是一个非常复杂的系统，具有语境敏感性、句法语义不对称、一词多义、模糊性等特征，欠规范现象比比皆是。对此，黑箱处理相当脆弱，几乎无能为力。例如，机器翻译界有一条著名的语句："The box was in the pen"。我们都知道"box"是盒子，"pen"有两个意思：一个是钢笔，一个是围栏。翻译这一语句，人们很容易给出正确翻译："盒子在围栏里"。然而，谷歌、百度、微软的机器翻译系统却将它翻译成"盒子在钢笔里"。为什么会这样？原因就在于，目前的机器翻译皆采用深度学习方法，直接依赖于大数据和概率统计。所以，要想得到正确的翻译，机器除了要知道"box"和"pen"的可能所指（习惯用法）之外，还应知道三点知识或常识：

（1）"in"是"一个小器具放在一个大器具里"。

（2）盒子的体积（通常情况下）小于围栏的体积，故而可以放在围栏里。

（3）盒子的体积（通常情况下）大于钢笔的体积，因此不能放在钢笔里。

令人遗憾的是，目前大数据驱动的机器翻译尚不具备这些最基本的人类知识或常识。机器在大规模的深度学习中根据概率获知"pen"常常译为"钢笔"，所以翻译系统便理所当然地把此处的"pen"误译为"钢笔"。假如翻译系统具备上述常识，它就能知道"盒子在钢笔里"这样的翻译是错误的。因为盒子只能装在围栏里，哪怕"围栏"这个词出现的概率再低，也只能译为"围栏"，而不能译为"钢笔"。

由上可知，大数据驱动的深度学习不过是在统计意义上将两个东西相关联，两者之间是

否具有逻辑关系，它却浑然不知。所以，计算机要想真正理解自然语言，仅仅依靠累积数据是远远不够的，还需要汇聚人类常识的大知识驱动。目前，大数据驱动的自然语言处理技术已经非常成熟了，而大知识驱动的自然语言处理技术才刚刚起步。虽然存在一些面向特定领域的专家知识库，但没有建构起面向全人类的大知识库，特别是常识库。以大数据和大知识为双轮驱动的自然语言处理，超越了传统经验主义和理性主义的二元竞争，是 AI 发展的必然趋势。

4. 文本生成

文本生成是自然语言处理中最有意思的任务之一。例如自动写诗，或自动作诗机、藏头诗生成器，目前支持五言绝句、七言绝句、五言律诗、七言律诗的自动生成（给定不超过 7 个字的开头内容自动续写）和藏头诗生成（给定不超过 8 个字的内容自动合成）。感兴趣的读者可以关注公众号 AINLP，如图 5.7 所示，看一下效果。

图 5.7　公众号 AINLP

文本生成是自然语言处理的一个重要的方向，例如，摘要生成要求机器阅读一篇文章后自动生成一段具有概括性质的内容，比如生成摘要或标题。与其他的应用不同，如机器对话、机器翻译等输入和输出文本的长度较为接近，文本摘要输入的文本长度往往远大于输出的文本长度，输入与输出的不对称也使得其较为特殊，因此诞生了一种提取式的方式——在原文中寻找重要的部分，将其复制并拼接成摘要，但提取的单词往往因为缺少连接词而不连续，而且无法产生原文中不存在但需要的新单词。因此，人们需要一种类似人类书写摘要的方法，先阅读文章并理解，再自己组织语言来编写摘要，摘要与原文意思接近且主旨明确。

机器像人一样使用自然语言进行表达和写作。依据输入的不同，文本生成技术主要包括数据到文本生成和文本到文本生成。数据到文本生成是指将包含键值对的数据转化为自然语言文本；文本到文本生成是对输入文本进行转化和处理从而产生新的文本。随着序列到序列（Sequence-to-Sequence，Seq2Seq）模型的成功，使用递归神经网络来阅读文章和生成题目成为可能。

除了以上四种研究方向外，还有信息过滤、信息检索等领域，这里由于篇幅所限，就不一一赘述了。

5.1.4　小结

虽然自然语言处理的相关研究比较抽象，但其最基础的研究还是对语法、句法和语义的研究，关注的核心是语言和文本。但因为语言现象所特有的不确定性及发展性，使得词、句、段落在不同的情景下都有其不同的含义，且伴随着文明的发展，新的词汇、语法层出不穷。这就要求计算机跳出对单个词、句的理解，增强对文本整体含义的把握能力，结合语境理解人类语言，真正地会思考、会创造，而非仅局限于功能性和使用性。这也将是自然语言处理的重要发展方向。

人工智能目前大致有两条发展路径，其一是以模型驱动的数据智能，其二是类脑科学与人工智能的研究。当前的主流发仍是数据智能，学习仍依赖大量的数据集合，具有很大的局

限性。自然语言处理要想取得突破性的进展，势必要从对人类大脑结构的研究入手，找出思维的奥秘，让机器真正听懂、读懂、看懂人类语言，而不再仅仅局限于对词、句的脱离语境的认知。其次，要重视学科交叉在自然语言处理应用落地时蕴含的巨大能量，促进自然语言处理＋行业的深度融合，让技术为企业带来更大的经济价值，为人们的日常生活增添更加便利性的新鲜元素。

5.2　智能语音

Siri是苹果公司在其产品上应用的一个语音助手，近日，国内外的很多朋友都在网上晒出了Siri的截图对话：

Can I borrow some money?

能给我借点钱不？

Ashleigh, you know that everything I have is yours.

借借借，你的是你的，我的也是你的！

Tell me a story.

给我讲个故事吧。

It was a dark and stormy night...no, that's not it.

那是一个月黑风高的夜晚……哎拉倒吧。

Where did I put my keys?

我把钥匙丢哪儿了？

It will probably be in the second-to-last place you look. Does that help?

可能在你找过的倒数第二个地方。问我特么的能有用？

☺☺☺

大家在戏弄Siri的同时，也别忘记Siri的其他神奇功能，比如利用地点设置提醒事项、智能呼叫、语音控制、球队比赛结果的获知等多种便利生活的功能。

从用户说话开始，到Siri的语音反馈，其实是经历了很多步骤的，如图5.8所示。

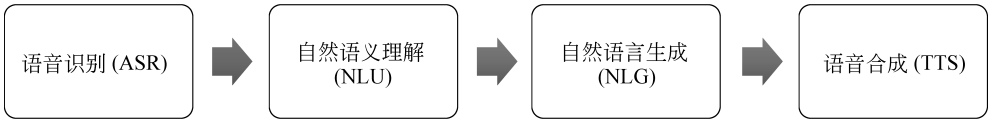


图5.8　Siri语音助手工作流程

第一步称为语音识别，就是将麦克风采集到的用户声音转化为文字的过程。

第二步称为自然语义理解，将用户说的话转化成机器能理解的话，例如，把转化成文字后的两句话“路挺滑，我差点儿没摔倒”和“路挺滑，我差点儿摔倒了”理解成同样的含义。

第三步称为自然语言生成，与自然语义理解相反，是将机器的语言转化人的语言，这个阶段输出的是文字。

最后一个阶段是语音合成，将文字合成声音并播放出来，并尽可能地模仿人类自然说话