第一章 绪论

1、中文信息处理（语言信息处理）

- 用计算机对自然语言在各个层次（语素、词、短语、句子、段落、篇章）上的各种表现形式（图像、声音、文本）的信息进行处理：输入、输出、转换、存储、压缩、检索、抽取和提炼。
- 语言信息处理往往是“用计算模型”而非仅仅是“用计算机”。

2、训练集、开发集、测试集

通常把经过人工标注或人工校对的实验用语料库划分为训练集、开发集、测试集三个集合，

训练集 用于提供模型参数的语料集

开发集 用于实验过程中检验和改进模型性能的语料集

测试集 用于实验结束后最终评判模型性能的语料集

例如，将实验用语料划分为 10 份，其中开发集和测试集各 1 份，其余 8 份作为训练集。

3、PRF

正确率：又称精确率（Precision），正确处理的实例个数占所处理的实例个数的比率。例如，人名识别的正确率等于正确识别的人名个数除以系统认为是人名的个数。

召回率：（Recall），正确处理的实例个数占应该处理的实例个数的比率。例如，人名识别的召回率等于正确识别的人名个数除以文本中实有的人名个数。

调和平均值：（F-measure），正确率和召回率的综合表示，$F=(\beta^2+1)PR/(\beta^2R+P)$， 通常取 $\beta=1$，则 $F=2PR/(R+P)$.

4、Topline：测试成绩的乐观估计。通常以人工处理的成绩作为 Topline。例如自动分词的顶线是人工分词。

Baseline：测试成绩的保守估计。通常用一种最简单可行的方法的成绩作为 BaseLine。例如自动分词的基线是最大匹配法分词。

5、封闭测试：运用从训练集里获取的数据（模型参数或规则）来测试训练集本身，目的是对模型的性能有初步了解。但是，封闭测试成绩可能主要反映模型对训练集的过度学习（学了太多琐碎的、依赖于罕见语境的数据）。

开放测试：运用从训练集里获取的数据来对测试集进行测试，目的是检验知识的覆盖能力。开放测试成绩通常低于封闭测试，但能够更真实地反映模型性能。

第二章 汉字处理

1、列举你所知道的字符编码集

等长码：GB2312、GBK、UTF-16

变长码：UTF-8、GB18030

- GB2312：6763 个汉字，不收繁体字
- Big5：港台，13053 个汉字，繁体字
- GBK：兼容 GB2312，含繁体字
- GB18030 ：1-4 字节编码方案，变长码
- Unicode 统一码：UTF8（变长）、UTF16（等长）
- ASCII

第三章 语言的表示形式

1、规则

规则是语言知识的经典表示形式，理性主义的方法通常叫做“基于规则”的方法。一般形式是 if…，then…，例如：

汉语语音规则：音节 → 声母+韵母+声调

词法规则：noun(复数) → noun(单数)+s

句法规则：S → NP+VP

规则库是用于处理某一类问题的规则的集合，例如词法规则库、句法规则库。

评价指标：

覆盖率：一条规则的条件被满足的次数与全部处理次数之比。

条件被满足，动作就会执行。但动作未必正确。该指标用来表示规则的使用频率，覆盖率高的规则，表达颗粒度大的知识。

正确率：一条规则获得正确处理结果的次数与该规则的条件被满足的次数之比。

该指标用来表示知识的质量。

好的规则应该是覆盖率和正确率都高

2、知识库

一种是指专家系统设计所应用的规则集合，包含规则所联系的事实及数据，它们的全体构成知识库。这种知识库是与具体的专家系统有关，不存在知识库的共享问题；另一种是指具有咨询性质的知识库，这种知识库是共享的，不是一家所独有的。从今后的发展来看，巨型知识库将会出现，还依赖于硬件及软件条件的发展。下一代计算机所应考虑的重要问题之一是知识库的设计，以知识库为背景的知识库公共管理系统机构设计。

3、电子词典

电子词典是语言知识的常见表现形式，通常存储于数据库，便于计算机存取。

狭义的“电子词典”专指词语知识库，每条记录是一个词或固定短语，有词性、词类、词义、读音、词频等字段。

广义的“电子词典”泛指语言知识库，其条目不限于词。包括计算机可读的字典、短语数据库、语素数据库、语音数据库、地名库、人名库、译名库等等

列举你所知道的几种电子词典：

- 北大：现代汉语语法信息词典：GBK

一部面向语言信息处理的大型电子词典，词典采用数据库文件格式，有总库和各词类分库，其中动词分库尤为详细，对于现代汉语的自动句法分析有重要价值。

- 梅家驹：同义词词林 【中英文双语知识网络。】
- 董振东：知网（Hownet） 【突出优点是词义代码，可据此计算词义之间的距离或相似度。】

知网中的“概念”相当于一个词义，概念是用一种知识描述语言来组织的一组“义原”，分为事件、实体、属性、属性值、动态角色等类别。

区别：规则库通常存储那些颗粒度较大的语言知识，电子词典通常存储那些颗粒度较小的语言知识。

4、语料库

语料库（Corpus）是为了语言研究按照一定的原则收集和组织的真实的自然语言作品（书面的和口头的）的集合。

一些重要的中文语料库：

1.北大现代汉语标注语料库（人民日报语料）
2.教育部现代汉语标注语料库（多种来源）
3.清华汉语树库（100 万字）
4.哈工大汉语树库（10 万字）
5.UPenn 树库（宾州树库，其中汉语树库 100 万字）
6.北语汉语中介语语料库
7.中科院计算所和自动化所英汉平行语料库（24 万句对）

5、语言模型

语言模型对自然语言及其运作机制的一种模拟工具。

狭义：用来度量语言符号串合法性的概率分布。

广义：语言信息处理时用来判别或预测的概率公式。

例如，机器翻译的统计模型中，包含一个目标语言模型 P(T)和一个翻译模型 P(S|T)，其中 P(T)就是狭义的语言模型。而下面这个关于机器翻译的公式就是广义的语言模型。

$$T = \arg\max_{T} P(T)P(S \mid T)$$

几种常见的语言模型：

贝叶斯分类器 NB
最大熵模型 ME
隐马尔科夫模型 HMM
条件随机场 CRF
决策树 DT　（以上都是概率模型）
word2vec　（向量空间模型）

比较：

①适用范围：单点分类可采用 NB 或 ME 等，序列标注可采用 HMM 或 CRF 等。
②分类性能：一般地说，ME 优于 NB，CRF 优于 HMM。但实际效果取决于许多因素，例如基元选择、训练集-测试集的划分、实际依赖关系的长短等等。
③时间效率：性能好的往往时间复杂度高。NB 优于 ME，HMM 优于 CRF。输出类别多时，CRF 难以胜任，例如音字转换。

CRF 的优势和应用

1.CRF 是一种用于序列标注的判别模型，其优点是较好地克服了输出独立性假设和马尔科夫性假设的局限性，能从上下文中任意地选择所需要的特征。

2.CRF 在基于字符的中文分词、命名实体识别、短语组块、基本名词短语识别等许多应用中都表现优异。

3.CRF 的主要缺点是参数规模过大，训练时间特别长。

第四章 自动分词

【三种基本分词思路】

1. 看成是一个字串切分问题：最大匹配法：采用贪心算法，不断地找出当前位置上的最长词。最大概率法：探查所有可能的词串，把概率最大的词串看做是最佳切分。
2. 看成是一个字符标注问题：CRF 训练和标注，字符角色的判别。
3. 看成是一个字符间隔标注问题：互信息标注，词界的判别。

1、中文自动分词的三大难题

1.未登录词：自动分词主要是根据底表来进行的，真实文本中存在大量的未见于底表的词语，它对自动分词正确率的影响最大.**eg**：考 研 的 时候；俄罗斯 总统 梅 德 韦 杰 夫

【应对方法】：基于构词知识的未登录词识别

2.分词歧义：根据底表，一个串可以切开也可以不切开（组合型歧义），或者可以切在这里也可以切在那里（交集型歧义），但从上下文来看，至少有一种切法是不正确的。

【应对方法】：组合型歧义——根据上下文，结合词性标注；交集型歧义——伪歧义：记录歧义字串和切分方式、词频，真歧义： 二元模型，引入上下文

3.分词不一致：上下文相同或相似情况下，一个串在分词语料库中有多种切法，也许几种切法都有道理，但应该保持一致。**eg**. 发展中国家/，发展中/国家/，发展/中/国家/

2、分词（切分）歧义

交集型歧义和组合型歧义

交集型歧义：如果一个字串有多种切分位置，并且每个字在不同切法中属于不同的词，那么这个字串称为交集型歧义字串，eg“这篇文章太平淡了”，使用户 满意；研究生命 的 起源

组合型歧义：如果一个字串是词，并且还可以看作是一个词串（至少包含两个词，每两个词互不交叠），那么这个字串就称为组合型歧义字串，eg：从马上跳下来;他 将来 我 校 讲 学

3、（自动分词的意义和价值）

自动分词（汉语特有的研究课题，中文信息处理的基础）

含义：把一个句子按照其中词的含义进行切分

过程：从信息处理的需要出发，按特定的规范，对汉语按分词单位进行划分的过程。

使用/计算机/将/字符串/自动/转换/为/词串

为什么要分词？

文本分析的第一步

中文信息处理

英语、日语

分词带来的帮助：

信息检索的预处理：提高查准率

语音合成的预处理：降低读音复杂性

汉字识别的后处理：提高识别正确率

语音识别的后处理：提高识别正确率
计算机辅助词典编撰：新词、新义项获取

4、自动分词的主要方法

机械切分：正向/逆向最大匹配法（速度快、实用）
简单的统计方法：最大概率（N 元模型）
当前主流：统计机器学习
隐马尔科夫（HMM）
最大熵（ME）
支持向量机（SVM）
条件随机场（CRF）
深度学习（RNN/LSTM）

5、最大匹配分词算法

分词思想：长度最小的词串（词最少）是最佳词串。
匹配：将汉字串跟词表中的词进行比较。
最大：长词优先，或称“最少分词法”。

- 词表：游戏、公司、天堂、任天堂
- 任天堂/游戏/公司

而不切分为：

- 任/天堂/游戏/公司

长词优先原则在绝大多数情况下是对的。

- *研究/生命/科学

几个概念：

①底表：词语的静态查找表，是关于“什么是词”的明确定义，不需要词频数据，也不必将单字词列入。
②最大词长：底表中最长词的长度，以字符为单位计算。
③候选词：从某位置开始截取的一个字符串，初始长度为 MIN(最大词长,剩余串长)。
候选词在底表中查找成功，便确定为词
候选词长度为 1 时不必查找，默认为词。

6、模拟自动分词

【匹配算法】：将汉字串跟词表中的词进行比较。
首先设定一个最大词长。某位置开始截取的一个字符串，作为候选词，它的初始长度为 MIN(最大词长,剩余串长)。对候选词在底表中进行查找。选词在底表中查找成功，便确定为词，找不到则将候选词末尾减一字，继续查找。候选词长度为 1 时不必查找，默认为词。

6、动态规划算法（过程）

使用动态规划算法的目的是减少计算量。
获取汉字串中全部候选词及其概率并转为费用；
对每个非首词的候选词，找出累计费用最小的前驱词作为最佳前驱词，该候选词的累计费用是它本身的费用加上最佳前驱词的累计费用；
对每个结尾的候选词，选择累计费用最小者作为最佳路径上的尾词；
从尾词开始，根据最佳前驱线索逆推最佳路径

第五章 实体识别

1、制定实体规范标准（操作）
2、实体属性
3、设计实体抽取算法

垃圾分类的对话系统需要用到哪些中文信息处理知识

首先收集大量垃圾数据，垃圾名称对应标注上垃圾类别（根据规定除可回收物、有害垃圾和湿垃圾都是干垃圾）、每个类别匹配上预先设定的随机回答、训练分类模型

信息抽取、命名实体识别：从用户的问题和对话中识别出垃圾实体，如“猪肉饺子是什么垃圾”识别出“猪肉饺子”、调用分类模型
语音处理：用户语音输入、系统语音输出，降低对用户的要求

论述题（人工智能、深度学习）

如何建立用药说明的实体标注体系

- 文献调研、数据分析确定规范初稿

1 确定标注基本体系，实体名称（BMES 四位标注体系-实体类别）+实体修饰成分（肯定与否定、程度）
2 标注文本范围确定，例如实验中我们小组的标注范围包含 C（药品名称）、F（通用名称）、G（成份）、V（注意事项）四列
3 实体的范围定义，确定实体类别（疾病、治疗、症状、药品主要成分、细菌病毒、与药发生作用的物质），强调非标注的范围（如器官与身体组织）
4 可纳入实体关系，例如具有某个治疗史则不可复用某种药品、某个症状可以确定某个疾病并确定对应治疗药品

- 抽样预标注
- 通过人机互助、医生协助、一致性评价（不同标注员是否会产生明显的差异）等过程，对基础标注规范进行迭代修订

方法：

1. 用标记来标注命名实体中各个词项的类别： B, C, L, U【14 级实验标注为 B/M/E/S 或者 BE/B/M/E】分别表示某类命名实体的首词、中词、尾词和独立词；
2. 为了避免非法的词标记序列，用二元转移概率来加以限制；
3. 定义特征空间；
4. 从训练语料中采集几个专用数据表；
5. 对训练语料中每个词标注其类别和特征，训练命名实体标注模型（CRF、最大熵、LSTM）；
6. 对测试语料中的每个词标注其特征，用对应的模型求出其类别。

4、IR 与 IE 之比较

1.输出不同：IR 文档列表； IE 事实信息
2.处理技术不同：IR 统计及关键词匹配；IE 模式匹配和统计分类
3.适用领域不同：IR 领域无关；IE 领域相关
4.信息检索与信息抽取又是互补的：
为了处理海量文本，IE 系统通常以 IR 系统（如文本过滤）的输出作为输入
IE 技术可用来提高 IR 系统的性能

5、信息抽取的研究意义（结合实例）

1.引入信息抽取技术，可帮助我们从 Web 中直接找出感兴趣的事实信息而不是一篇篇文档。

2.数据挖掘是从大量的原始数据中获取知识，信息抽取技术可为数据挖掘提供大量的原始数据。

3.直接用于各种文本信息的处理，例如从期刊中抽取科学论文的各种数据，从古典文献中抽取历史事件的各种数据，以建立相应的信息检索系统。

4.信息抽取是一种基于内容的计算，一种比较初级的自然语言理解，比较适合于当前的技术水平。

第六章 词性标注

1、两种关于词类体系的理论

两种关于词类体系的理论

- 词类多功能说
  - 一种词类可以有多种功能，例如动词、形容词和名词，其词类不依出现位置的变化而变化，动词、形容词都可以作谓语、补语，也都可以做主宾语
  - 词的语法功能是潜在的，每次只实现它的一种功能，但其它功能并非就消失了
  - 汉语的词在实现其语法功能时没有词形变化，因此没有根据说不同位置上的词属于不同词类

- 依句辨品说（“依句辨品，离句无品”）
  - 应依据词在句子中所实现的功能来确定其词类。动词、形容词在谓语位置上是动词，在主宾语位置上则是名词。
  - 这本书出版了　　动词
  - 出版了这本书　　动词
  - 辞书出版上取得了很大成绩　　？名词
  - 这本书的出版是有原因的　　名词
  - 这本书还没列入出版计划　　？形容词
  - 出版的可能性很大　　？形容词
  - 这本书的不出版是有原因的　　？？

2、词性标注的三个主要问题

- 词类体系和词性标记集：词类体系是理论问题，词性标记集是在词类体系基础上建立的，是技术实现问题。（一个词类用什么作为标记，需考虑处理是否方便。）
- 兼类词的词性消歧：一个词条包含多个词型，且这几个词型的语法功能总和有所区别，这个词条或者跟这个词条写法相同的词例叫做兼类词。（这里说的兼类词包括同形词。）
- 未登录词的词性猜测：猜测它的语法功能总和，而非猜测它正实现的语法功能。（后者应由自动句法分析来承担。）

3、词性标注的意义

1.确定词的语法功能，为自动句法分析打基础:词的语法功能，就是一个词能充当什么句法成分，例如做主语、宾语、谓语、定语、状语和补语等等。

2.在词性标注语料库中检索句法结构:许多结构是以特定的词类为标志的，例如介词结构、“把”字结构、方位结构等等。

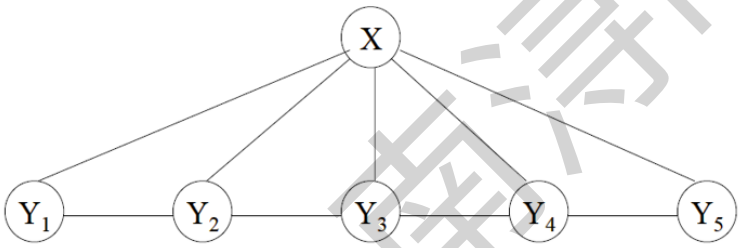
3.为多音字消歧和多义词消歧提供支持:多音字经词性标注后减少了读音。多义词经词性标注后减少了歧义。

4、CRF 词性训练过程

使用CRFs进行词性标注

- CRFs模型
- 摆脱了独立性假设，更适应序列化标注任务

Suppose $P(Y_v \mid X, \text{all other } Y) = P(Y_v \mid X, \text{neighbors}(Y_v))$
then X with Y is a **conditional** random field



- $P(Y_3 \mid X, \text{all other } Y) = P(Y_3 \mid X, Y_2, Y_4)$
- Think of X as observations and Y as labels

CRFs模型包的使用

- CRF++: Yet Another CRF toolkit v0.54
  - http://crfpp.sourceforge.net/
  - 语料格式
  - 模板格式
  - 训练用法
  - 自动标注
- 注意
  - 对硬件要求高，在Win32环境下，只能使用2G内存
  - 上下文窗口，最大为正负4
  - 输入是单个语料文件，为gbk或utf-8编码（char作为字符串类型）

第七章 词义消歧

什么是词义消歧

- 词义消歧（Word Sense Disambiguation）：对于文本中的多义词形，辨别它是哪一个意思，叫做词义消歧。
- 歧义词：包括同形词、多义词。通常根据词典、义类词典或其他资源来确定它有哪几个意思。
- 词义消歧的依据：歧义词所在的上下文，词典释义，双语对齐文本。
- 词义消岐通常是对特定的多义词形的处理。词义标注则是对文本中所有的多义词形的处理。后者难度更大。

1、回指、共指

指代消解：解决多个指称对应一个实体的问题

共指消解：

确定名词性成分之间的共指关系的过程，从而合并从多处（篇章内、之间）不同描述中得到的语义结构。例如，人名、地名等实体都存在异名同指和同名异指现象，需要进行共指消解。

回指（anaphora）：一个语言成分指代上文中出现过的语言成分。用来指示的语言成分叫做指示语，通常包含指代词，被指示的语言成分叫做先行语。例如：王先生是江苏人，他有一个儿子和一个女儿 。

先行语与指示语之间没有等价关系，依赖于上下文。

共指（coreference）：一些名词短语（包括命名实体）虽然形式不同，但都指同一个实体，是同一个实体的几个提及（mention）。前例之外，又如：

三月，公及邾儀父盟于蔑 （左传隐公元年）

三月，公及邾婁儀父盟于昧（公羊传隐公元年）【8/9 考一个】

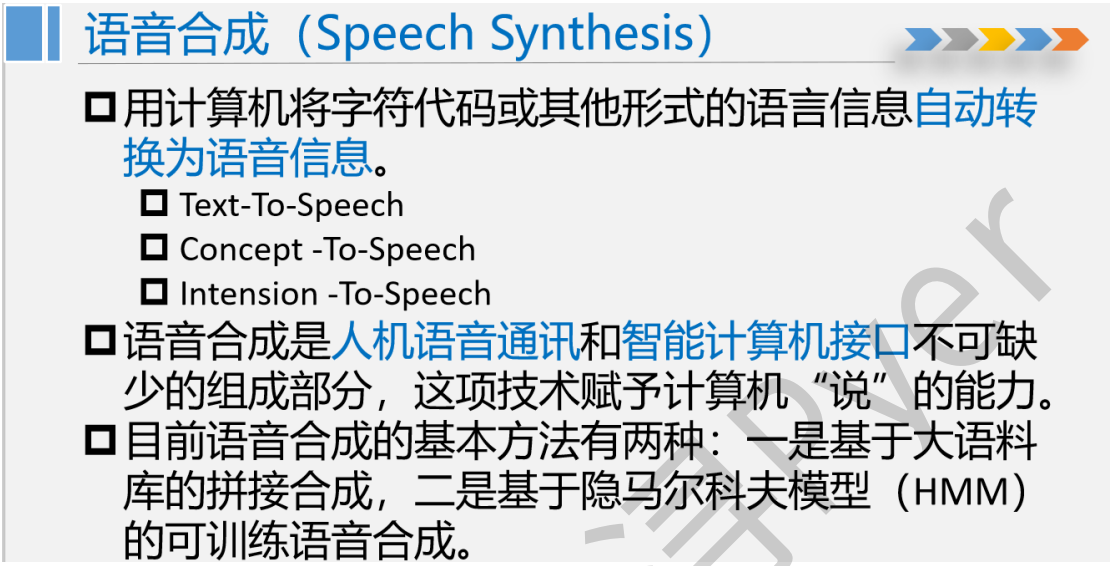
共指成分之间有等价关系，且独立于上下文。

第八章

语言信息处理的应用研究课题

汉字识别、汉语语音识别、汉语语音合成、汉语信息检索、汉外机器翻译

1、语音合成



2、机器翻译

机器翻译，又称为自动翻译，是利用计算机将一种自然语言(源语言)转换为另一种自然语言(目标语言)的过程。它是计算语言学的一个分支，是人工智能的终极目标之一，具有重要的科学研究价值。

发展历程：传统、机器学习、深度学习

机器翻译的研究经历了基于规则的方法、基于实例的方法、基于统计的方法、基于神经网络的方法四个阶段的发展。

（理解吧还是，当做看一个故事）

在机器翻译研究的早期，主要 使用基于规则的方法。机器翻译系统根据语言专家编写的翻译规则进行翻译，这是一个机械式的过程。基于规则的方法受限于人工编写的规则的质量和数量，编写规则非常费时费力，且翻译规则无法用于不同的语言对之间。同时，规则数量增多，互相冲突的规则也随之增多，难以覆盖人类语言的全部情况，这也是机器翻译系统 的瓶颈。

20 世纪 90 年代，基于统计的机器翻译方法被提出，随后迅速成了机器翻译研究的主流方法。统计机器翻译使用双语平行语料库（即同时包含源语言和与其互为译文的目标语言文本的语料库，作为训练数据。

近年来随着基于神经网络的方法被引入机器翻译领域，机器翻译的性能得到了大幅提高。根据谷歌机器翻译团队发布的信息，谷 歌翻译于 2016 年 9 月上线中英神经网络模型，截至 2017 年 5 月，已经支持 41 对双语翻译模块，超过 50% 的翻译流量已经由神经网络模型提供。

机器翻译一个新的趋势是正在“实用化”，被应用到生活场景中。过去机器翻译像是一个“更智能的词典”，帮助人们阅读外文网页内容。现在随着语音和图像识别技术的进步，机器翻译可以更多地与生活场景结合。比如人们出国时可用百度翻译了解菜单、店名、商品信息，看美剧时可以用电脑进行字幕翻译，通过拍照直接翻译出一朵花的名字，再比如开头提到的机器人进行多语翻译采访等等。

基于规则的机器翻译

机器翻译的源头，可以追溯至 1949 年，资讯理论研究者 Warren Weave 正式提出了机器翻译的概念。五年后，也就是 1954 年，IBM 与美国乔治敦大学合作公布了世界上第一台翻译机 IBM-701。它能够将俄语翻译为英文，别看它有巨大的身躯，事实上它里面只内建了 6 条文法规则，以及 250 个单字。但即使如此，这仍是技术的重大突破，那时人类开始觉得应该很快就能将语言的高墙打破。但其实它并未提到翻译所用到的例子是经过了精心的挑选和测试，并排除了任何歧义。这个系统实际上无外乎形同一本常用语手册。然而，包括加拿大、德国、法国、尤其是日本，各国间就此展开了竞争，所有人都加入了机器翻译的比拼。但是由于规则太复杂，太费语言学家，老头子顶不住，发展停滞了。

基于实例的机器翻译

在全世界都陷入机器翻译低潮期，却有一个国家对于机器翻译有着强大的执念，那就是日本。日本人的英文能力差举世皆知，也因此对机器翻译有强烈的刚性需求。日本京都大学的长尾真教授提出了基于实例的机器翻译，也就是别再去想让机器从无到有来翻译，我们只要存上足够多的例句，即使遇到不完全匹配的句子，我们也可以比对例句，只要替换不一样的词的翻译就可以。这种天真的想法当然没有比基于规则的机器翻译高明多少，所以并未引起风潮。这个方法虽然不算是一次彻底的变革，但显然是向前迈进了一大步。仅在 5 年后，革命性的发明——统计型机器翻译出现了。

基于统计的机器翻译

统计模型的思路是把翻译当成机率问题。原则上是需要利用平行语料，然后逐字进行统计。例如，机器虽然不知道“知识”的英文是什么，但是在大多数的语料统计后，会发现只要有知识出现的句子，对应的英文例句就会出现“Knowledge”这个字。如此一来，即使不用人工维护词典与文法规则，也能让机器理解单词的意思。这种机器翻译方法使用的文本越多，翻译效果就越佳。事实上这种翻译方法已经相当不错，后续很多公司的翻译软件都是基于统计的翻译方式。

神经网络机器翻译

到了 2014 年，机器翻译迎来了史上最革命的改变——“深度学习”来了！通过提取语言句子的特征来进行翻译，尤其是 RNN 神经网络（该网络可以记住之前的结果，对文本来说即为之前的单词）广泛应用。工作原理大概是一个网络用来特征提取编码，另一个神经网络用来解码回归原本的语言文本。