西安交通大学
XI'AN JIAOTONG UNIVERSITY

人工智能技术导论

第六章 自然语言处理

内 容

6.1 自然语言处理概述

6.2 机器翻译

6.3 自然语言人机交互

6.4 智能问答

西安交通大學

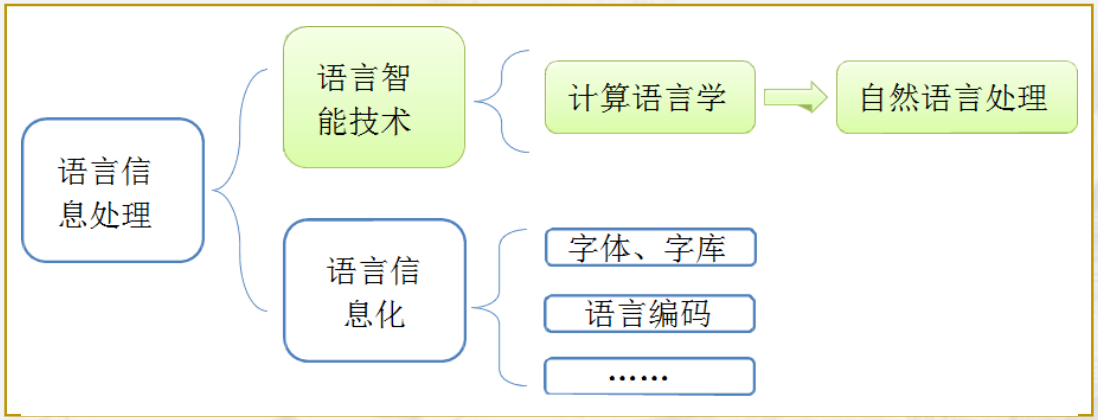
XI'AN JIAOTONG UNIVERSITY

6.1 自然语言处理概述

语言智能

只有人类具有语言智能能力。但人类并没有专门感知语言的器官。人类通过视觉、听觉感知承载语言的图像和声音信号，再经过大脑加工和抽象后，才能形成语言信息。因此语言不是一种感知信号，而是感知信号的经大脑处理后的某种抽象表示。所以，语言智能属于“认知智能”的研究范畴，是认知智能研究的核心问题之一。

在语言智能以及相关的研究领域中，有若干研究方向和术语，常见的有“自然语言处理”、“语言信息处理”、“计算语言学”。三者之间有很多交叉关联。



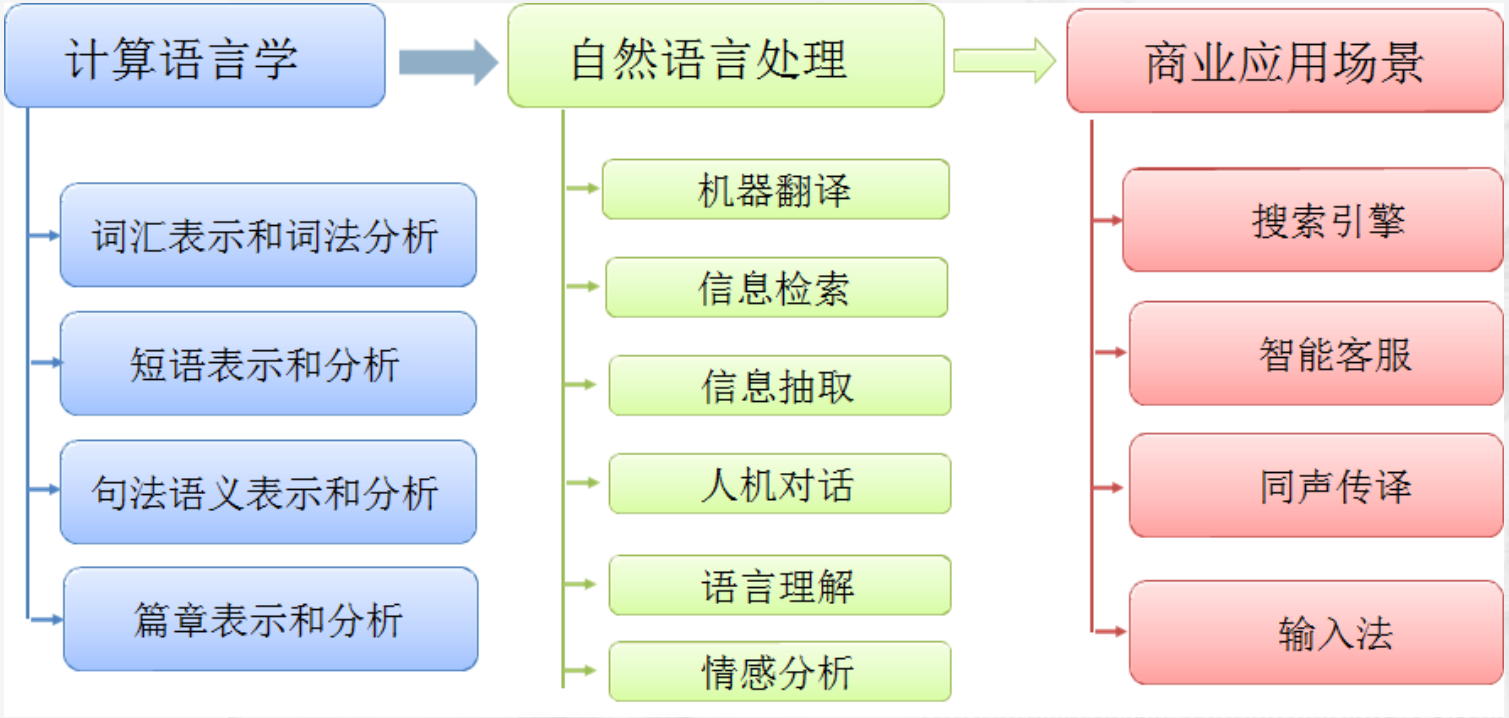
计算语言学（ Computational Linguistics ）指的是这样一门学科，它通过建立形式化的数学模型，来分析、处理自然语言，并在计算机上用程序来实现分析和处理的过程，从而达到以机器来模拟人的部分乃至全部语言能力的目的.人与计算机之间交流信息要使用计算机语言。

电脑每做的一次动作，一个步骤，实际上都是执行已经用计算机语言编好的程序。程序是计算机要执行的指令的集合，而程序全部都是用我们所掌握的语言来编写的。人们要控制计算机，利用计算机来解决问题，就一定要通过计算机语言向计算机发出命令。我们把编写程序的过程叫做程序设计，而计算机语言相应地称为程序设计语言 。计算机语言都可以用来控制计算机来解决一些实际问题。这些问题可以是数值计算问题，其操作对象就是一些由符号构成的符号串；也可以是非数值计算问题如声音、图像处理问题，其操作对象就是声音和图像等。我们应知道各种计算机语言都不是万能的，每种计算机语言都有自己的特点、优势及运行环境，有自己的应用和操作对象

语言智能

在了解语言智能研究的范畴以后，大致可以认为语言智能技术研究两方面内容：

- 建立语言的形式化模型，并在计算机上实现语言的表示和分析；
- 使用这样的形式化模型，用计算机解决各种与自然语言相关的实际问题，从而让计算机能够模拟人类的语言能力.



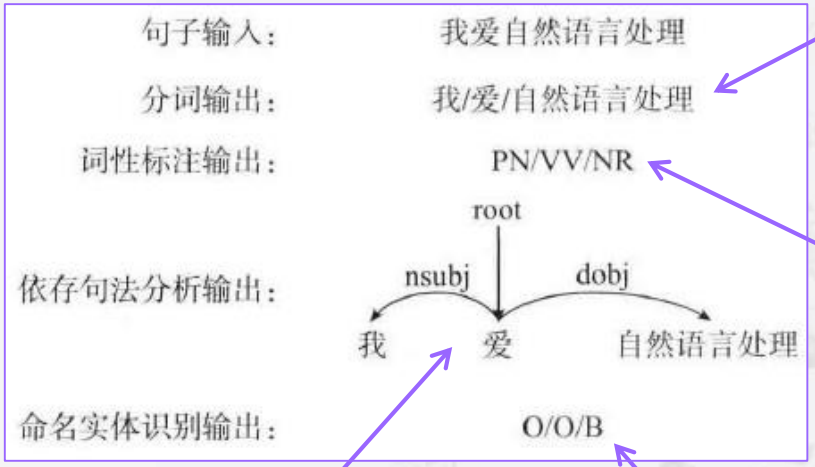
语言智能的研究和应用范畴

自然语言处理NLP

NLP是人工智能的一个分支，用于分析、理解和生成自然语言，以方便人和计算机设备进行交流，以及人与人之间的交流。

NLP中四个最基本的任务——分词、词性标注、依存句法分析和命名实体识别。

实例：



① 分词模块负责将输入汉字序列切分成单词序列，在该例子中对应的输出是“我/爱/自然语言处理”。该模块是NLP中最底层和最基础的任务，其输出直接影响后续模块。

② 词性标注模块负责为分词结果中的每个单词标注一个词性，如名词、动词和形容词等。在该例子中对应的输出是“PN/VV/NR”。

- PN 表示第一个单词“我”，对应词性是代词；
- VV 表示第二个单词“爱”，对应词性是动词；
- NR 表示第三个单词“自然语言处理”，对应的词性是专有名词。

③ 依存句法分析负责预测句子中单词与单词间的依存关系，并用树状结构来表示整句的句法结构。

- root 表示单词“爱”是整句对应依存句法树的根节点；
- 依存关系nsubj表示单词“我”是单词“爱”对应的主语；
- 依存关系dobj 表示单词“自然语言处理”是单词“爱”对应的宾语。

④ 命名实体识别负责从文本中识别出具有特定意义的实体，如人名、地名、机构名、专有名词等。在该例子中对应的输出是“O/O/B”。

- O表示前两个单词“我”和“爱”并不代表任何命名实体；
- B表示第三个单词“自然语言处理”是一个命名实体。

自然语言处理NLP

发展历程

第一阶段（60—80年代）：基于**规则**来建立词汇、句法语义分析、 问答、聊天和机器翻译系统。

好处是规则可以利用人类的内省知识，不依赖数据，可以快速起步；问题是覆盖面不足，像个玩具系统，规则管理和可扩展一直没有解决。

第二阶段（90年代开始）：基于统计的**机器学习**开始流行，很多NLP开始用基于统计的方法来做。

主要思路是利用带标注的数据，基于人工定义的特征建立机器学习系统，并利用数据经过学习确定机器学习系统的参数。运行时利用这些学习得到的参数，对输入数据进行解码得到输出。机器翻译、搜索引擎都是利用统计方法获得了成功。

第三阶段（2008年之后）：深度学习开始在语音和图像发挥威力。随之，NLP研究者开始把目光转向深度学习。

先是把深度学习用于特征计算或者建立一个新的特征，然后在原有的统计学习框架下体验效果。比如，搜索引擎加入了深度学习的检索词和文档的相似度计算，以提升搜索的相关度。自2014年以来，人们尝试直接通过深度学习建模，进行端对端的训练。目前已在机器翻译、问答、阅读理解等领域取得了进展，出现了深度学习的热潮。

自然语言处理NLP

发展历程

第三阶段（2008年之后）：深度学习

① 神经网络的端对端训练使自然语言处理技术不需要人工进行特征抽取，只要准备好足够的标注数据(如机器翻译的双语对照语料)，利用神经网络就可以得到一一个现阶段最好的模型；

② 词嵌人（word embedding）的思想使得词汇、短语、句子乃至篇章的表达可以在大规模语料上进行训练，得到一个在多维语义空间上的表达，使得词汇之间、短语之间、句子之间乃至篇章之间的语义距离可以计算；

③ 基于神经网络训练的语言模型可以更加精准地预测下一个词或一个句子的出现概率；

④ 循环神经网络（RNN、LSTM、GRU）可以对一个不定长的句子进行编码，描述句子的信息；

⑤ 编码-解码（encoder-decoder）技术可以实现一个句子到另外一个句子的变换，这个技术是神经机器翻译、对话生成、问答、转述的核心技术；

⑥ 强化学习技术使得自然语言系统可以通过用户或者环境的反馈调整神经网络各级的参数，从而改进系统的性能。

神经机器翻译

神经机器翻译是模拟人脑的翻译过程。人在翻译的时候，首先是理解这句话，然后在脑海里形成对这句话的语义表示，最后再把这个语义表示转化为另一种语言。神经机器翻译有两个模块：

- 一个是编码模块，把输入的源语言句子变成一个中间的语义表示，用一系列的机器内部状态来代表；
- 另一个模块是解码模块，根据语义分析的结果逐词生成目标语言。

一个重要的研究问题就是数据问题，神经机器翻译**依赖于双语对照的大规模数据集**来进行端到端的训练神经网络参数，这涉及很多语言对和很多的垂直领域。而在某些领域并**没有那么多的数据**，只有少量的双语数据和大量的单语数据，所以如何进行半监督或者无监督训练来提升神经机器翻译的性能成为本领域的研究焦点。

神经机器翻译在这几年发展得非常迅速，现在神经机器翻译已经取代统计机器翻译，成为机器翻译的主流技术。统计数据表明，在一些传统的统计机器翻译难以完成的任务上，神经机器翻译的性能远远超过了统计机器翻译，而且跟人的标准答案非常接近甚至说是相仿的水平。

中文：**我就是这样评价**今年的女子奥运体操队的，原因不止一个。

统计机器翻译：**I said of** this year's women's Olympic gymnastics team, because more than one.

神经机器翻译：**This is how I evaluate** this year's women's Olympic gymnastics team, for more than one reason.

人工翻译：**That's what I call** this year's women's Olympic gymnastics team and for more reasons than one.

自然语言处理概述

智能人机交互

智能人机交互是指利用自然语言实现人与机器的自然交流。——“对话即平台”

- 大家都已经习惯社交手段（如微信）与他人聊天的过程。我们希望将这种通过自然语言交流的过程呈现在当令的人机交互中，而语音交流的背后就是对话平台。
- 在于现在大家面对的设备有的屏幕很小甚至没有屏幕，所以通过语音的交互更为自然和直观。

因此，我们需要对话式的自然语言交流，例如借助语音助手来完成。而语音助手又可以调用很多Bot（对话机器人）来完成一些具体功能，比如买咖啡、买车票等需求，每个需求都有可能是一个小Bot，必须有人去做这个Bot。许多公司希望将自己的能力通过开做平台释放出来，让全世界的开发者甚至普通学生都能开发出自己喜欢的Bot，形成一个生态的平台和环境。

如何从人出发，通过智能助理，再通过Bot体现这一生态呢?

- 第一个是面向任务的对话系统，比如微软的小娜，通过手机和智能设备介人，让人与电脑进行交流：人发布命令，小娜理解并完成任务。同时，小娜作为你的贴身处理，也理解你的性格特点、喜好、习惯，然后主动给你一些贴心提示。
- 第二个是聊天机器人，比如微软的小冰，主要负责闲聊。无论是小冰这种闲聊，还是小娜这种注重任务执行的技术，其实背后单元处理引擎无外乎三层技术：

聊天机器人要理解人的意图，产生比较符合人的想法以及符合当前上下文的回复，再根据人与机器各自的回复将话题进行下去。基于当前的输人信息，再加上对话的情感以及用户的画像，经过一个类似于神经机器翻译的解码模型生成回复，可以达到上下文相关、领域相关、话题有关且是针对用户特点的个性化回复。

① **通用聊天。**需要掌握沟通技巧、通用聊天数据、主题聊天数据，还要知道用户画像，投其所好；

② **信息服务和问答。**需要搜索的能力、问答的能力，还需要对常见问题表进行收集、整理和搜索，从知识图表、文档和图表中找出相应信息并回答问题，我们统称为Info Bot；

③ **面向特定任务的对话能力。**如买咖啡、定花、买火车票这些任务是固定的，状态也是固定的，状态转移也是清晰的，那么就可以用Bot一个个实现。它用到的技术是对用户意图的理解、对话的管理、领域知识、对话图谱等。

阅读理解

自然语言理解的一个重要研究课题是阅读理解。阅读理解就是让电脑看一篇文章，并针对文章问一些问题，看电脑能不能回答出来。

斯坦福大学曾做过一个比较有名的实验，就是使用维基百科的文章提出5个问题，由人把答案做出来，然后把数据分成训练集和测试集。训练集是公开的，用来训练阅读理解系统，而测试集不公开，个人把训练结果上传给斯坦福大学，斯坦福大学在其云端运行，再把结果报在网站上，这也避免了一些人对测试集做手脚。

阅读理解技术自2016年9月前后发布，就引起了很多研究单位的关注，大概有二三十家单位都在做这样的研究。一开始的水平都不是很高， 以100分为例，人的水平是82.3分，机器的水平只有74分，相差甚远，后来通过类似于开源社区模式的不断改进，性能得以逐步提高。

2018年在阅读理解领域出现了一个备受关注的问题，就是如何才能做到超越人的标注水平。当时微软、阿里巴巴、科大讯飞和哈尔滨工业大学的系统都超越了人工的标注水平，这标志着阅读理解技术进入了一个新的阶段。 而这几个系统都来自中国，也体现了中国在自然语言处理领域的进步。

阅读理解框架首先要得到每个词的语义表示，再得到**每个句子的语义表示**，这可以用循环神经网络RNN来实现，然后用特定路径来**找出潜在答案**，基于这个答案再**筛选出最优答案**，最后确定这个答案的边界。在做阅读理解时用到了外部知识，可以用大规模的语料来**训练外部知识**，通过将外部知识训练的RNN模型加人原来端到端的训练结果中，可以大幅度提高阅读理解的能力。

自然语言处理概述

机器创作

除了可以做到性的东西，机器可以做些创造性的东西吗？

- 2005年微软研究院研发成功微软对联系统。用户出上联，电脑对出下联和横批，语句非学称在此基础上，进一步开发了猜字谜的智能系统。 在字谜游戏里， 用户给出谜面，让乐统猜字；或系统给出谜面，让用户猜字。此后，关于创作绝句、律诗、唐诗来词等的研究随之兴起。
- 2017年，微软研究院开发了电脑写诗、作词、谱曲系统。中央电视台《机智过人》节目就曾播出过微软的电脑作词谱曲与人类选手进行词曲创作比拼的节目。这件事说明，如果有大数据，那么机器学习或者深度学习就可以模拟人类的创造智能造出一些作品来;也可以与专家合作，帮助专家产生更好的想法。这在以前是难以想象的，自然语言的研究人员从来没有想到自然语言还可以延伸到音乐上去（其实音乐也是一一种语言， 自然语言的所有技术都可以应用到音乐上去，这需要想象力）。

随着大数据、深度学习、计算能力、场景等的推动，预计未来NLP会进人爆发式的发展阶段，从NLP基础技术到核心技术再到NLP+的应用都会取得巨大进步。

- 口语翻译会完全普及，拿起手机一口语识别一翻译、语音合成实现一气呵成的体验；
- 自然语言会话(包括聊天、问答、对话)在典型的场景下完全达到实用；
- 自动写诗、新闻小说流行歌曲开始流行。

自然语言尤其是会话的发展大大推动语音助手、物联网、智能硬件和智能家居的实用化，这些基本能力的提升会带动各行各业如教育、医疗、法律等领域的生产流程。人类的生活发生重大变化，NLP也会惠及更多的人。然而，还有很多需要解决的问题。

- 个性化服务，无论是翻译、对话还是语音助手，都要避免千人一面的结果，要实现内容个性化、风格个性化、操作个性化，要记忆用户的习惯，避免重复提问。
- 目前基于深度学习的机制都是端对端训练练，不能解释、无法分析机理，需要进一步发展深度学习的可理解和可视化，可跟踪错误分析原因。很多领域有人类知识(如翻译的语言学知识、客服的专家知识),如何把数据驱动的深度学习与知识相互结合以提高学习效率和学习质量，是一个值得重视的课题。

此外，在一个领域学习的自然语言处理模型（如翻译系统）如何通过迁移学习来很好地处理另个领域? 还有如何巧妙运用无标注数据来有效缓解对标注的压力？等都是需要持续研究的方向。

自然语言处理概述

机器创作

ChatGPT

ChatGPT是由人工智能研究实验室OpenAI在2022年11月30日发布的全新聊天机器人模型，一款人工智能技术驱动的自然语言处理工具。它能够通过学习和理解人类的语言来进行对话，还能根据聊天的上下文进行互动，真正像人类一样来聊天交流，甚至能完成撰写邮件、视频脚本、文案、翻译、代码等任务。

ChatGPT使用了**Transformer神经网络架构，也是GPT-4.0架构**，这是一种用于处理序列数据的模型，拥有语言理解和文本生成能力，尤其是它会通过连接大量的语料库来训练模型，这些语料库包含了真实世界中的对话，使得其具备上知天文下知地理。

- ChatGPT具有同类产品具备的一些特性，例如对话能力，能够在同一个会话期间内回答上下文相关的后续问题。然而，其在短时间内引爆全球的原因在于，ChatGPT不仅能流畅地与用户对话，甚至能写诗、撰文、编码。
- ChatGPT还采用了注重道德水平的训练方式，按照预先设计的道德准则，对不怀好意的提问和请求“说不”。一旦发现用户给出的文字提示里面含有恶意，包括但不限于暴力、歧视、犯罪等意图，都会拒绝提供有效答案。

ChatGPT受到关注的重要原因是引入新技术RLHF（Reinforcement Learning with Human Feedback，基于人类反馈的强化学习）。RLHF 解决了生成模型的一个核心问题，即如何让人工智能模型的产出和人类的常识、认知、需求、价值观保持一致。ChatGPT是AIGC（AI- Generated Content，人工智能生成内容）技术进展的成果。该模型能够促进利用人工智能进行内容创作、提升内容生产效率与丰富度。

ChatGPT 的使用上还有局限性，模型仍有优化空间。ChatGPT模型的能力上限是由奖励模型决定，该模型需要巨量的语料来拟合真实世界，对标注员的工作量以及综合素质要求较高。ChatGPT可能会出现创造不存在的知识，或者主观猜测提问者的意图等问题，模型的优化将是一个持续的过程。若AI技术迭代不及预期，NLP模型优化受限，则相关产业发展进度会受到影响。此外，ChatGPT盈利模式尚处于探索阶段，后续商业化落地进展有待观察。

自然语言处理概述

机器创作

DeepSeek

DeepSeek是杭州深度求索公司发布的一系列在知识类任务上表现出色的人工智能模型，被称为“AI界的拼多多”。

DeepSeek的主要功能包括自然语言查询处理、代码生成、API和Web服务等。其核心特点在于其卓越的多语言编程能力、强大的上下文支持、高效的推理速度和开源策略。

DeepSeek模型采用了混合专家（MoE）架构，通过动态选择最合适的专家进行计算，提高了计算效率。同时，DeepSeek还通过算法和工程上的创新，实现了生成吐字速度的大幅提升，为用户带来更加迅速流畅的使用体验。此外，DeepSeek还新增了“作者朗读音色”功能，利用深度学习中的合成语音技术，能够生成与特定作者相似的朗读声音，进一步丰富了其应用场景。

- DeepSeek模型以其高质量编码服务而著称，不仅提供了通用的开源模型，还专门针对编码任务开发了名为DeepSeek Coder的模型。DeepSeek的起源可以追溯到人工智能和机器学习技术的快速发展时期，DeepSeek旨在利用先进的自然语言处理和机器学习技术，为用户提供高质量的编码服务。
- DeepSeek的应用领域非常广泛，主要涵盖了软件开发、数据分析、自然语言处理等多个方面。在软件开发领域，DeepSeek的编码服务能够帮助开发者更快速地完成代码编写和调试工作，提高开发效率和质量。

与传统的编码方式相比，DeepSeek的编码服务更加高效和智能化，能够自动生成高质量的代码，大大提高了编码效率和质量。与市场上的其他编码服务产品相比，DeepSeek允许用户根据自己的需求进行定制和扩展，能够更好地满足用户的实际需求。在性能方面，DeepSeek-V3在多项基准测试中表现出色，超越了包括Meta公司的Llama-3.1-405B和阿里云的Qwen 2.5-72B等一众领先开源模型，甚至在部分测试中超越了OpenAI的闭源模型GPT-4o。此外，DeepSeek还以其低推理成本在业界获得了“AI界的拼多多”的称号。

随着人工智能和机器学习技术的不断进步，DeepSeek的未来发展趋势充满了无限可能，DeepSeek有望成为人工智能和编码领域内的一颗璀璨明珠。

西安交通大学

XI'AN JIAOTONG UNIVERSITY

6.2 机器翻译

机器翻译

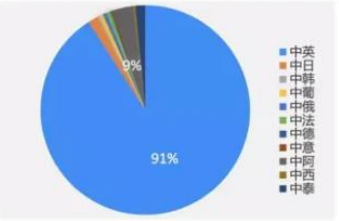
机器翻译，又称为自动翻译，是利用计算机**将一种自然语言（源语言）转换为另一种自然语言（目标语言）**的过程。它是计算语言学的一个分支，是人工智能的终极目标之一，具有重要的科学研究价值，但面临如下挑战：

① 译文选择。在翻译一个句子的时候，会面临很多选词的问题，因为语言中一词多义的现象比较普遍。比如源句子中的『看』，可以翻译成『look』、『watch』 『read 』和 『see』等词，如果不考虑后面的宾语『书』的话，这几个译文都对。在这个句子中，只有系统知道『看』的宾语『书』，才能做出正确的译文选择，把『看』翻译为『read』，『read a book』。

② 词语顺序。由于文化及语言发展上的差异，我们在表述的时候，有时候先说这样一个成份，后面说另外一个成份，但是，在另外一种语言中，这些语言成分的顺序可能是完全相反的。比如『在周日』，这样一个时间状语在英语中习惯上放在句子后面。再比如，像中文和日文的翻译，中文的句法是『主谓宾』，而日文的句法是『主宾谓』，日文把动词放在句子最后。比如中文说『我吃饭』，那么日语呢就会说『我饭吃』。当句子变长时，语序调整会更加复杂。

③ 数据稀疏。据不完全统计，现在人类的语言大约有超过五千种。现在的机器翻译技术大部分都是基于大数据的，只有在大量的数据上训练才能获得一个比较好的效果。而实际上，语言数量的分布非常不均匀的。右边的饼图显示了中文相关语言的一个分布情况，大家可以看到，百分之九十以上的都是中文和英文的双语句对，中文和其他语言的资源呢，是非常少的。在非常少的数据上，想训练一个好的系统是非常困难的。

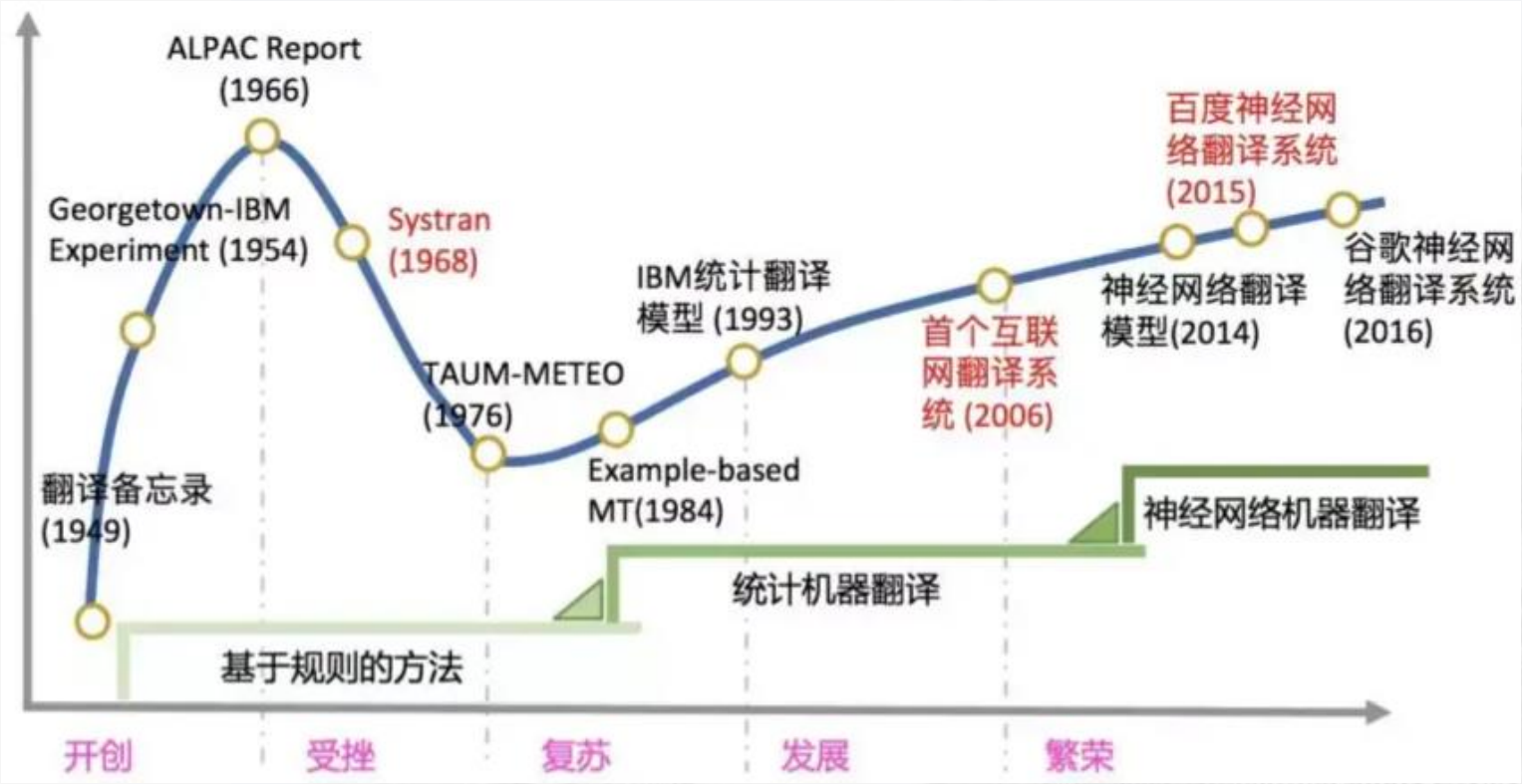
人类语言超过5000种



中文相关主要语种双语资源分布

机器翻译

机器翻译的研究历史可以追溯到20世纪三四十年代。20世纪30年代初，法国科学家G.B.阿尔楚尼提出了用机器来进行翻译的想法。1933年，苏联发明家П.П.特罗扬斯基设计了把一种语言翻译成另一种语言的机器，并在同年9月5日登记了他的发明；但是，由于30年代技术水平还很低，他的翻译机没有制成。1946年，第一台现代电子计算机ENIAC诞生，随后不久，信息论的先驱、美国科学家W. Weaver和英国工程师A. D. Booth在讨论电子计算机的应用范围时，于1947年提出了利用计算机进行语言自动翻译的想法。1949年，W. Weaver 发表《翻译备忘录》，正式提出机器翻译的思想。

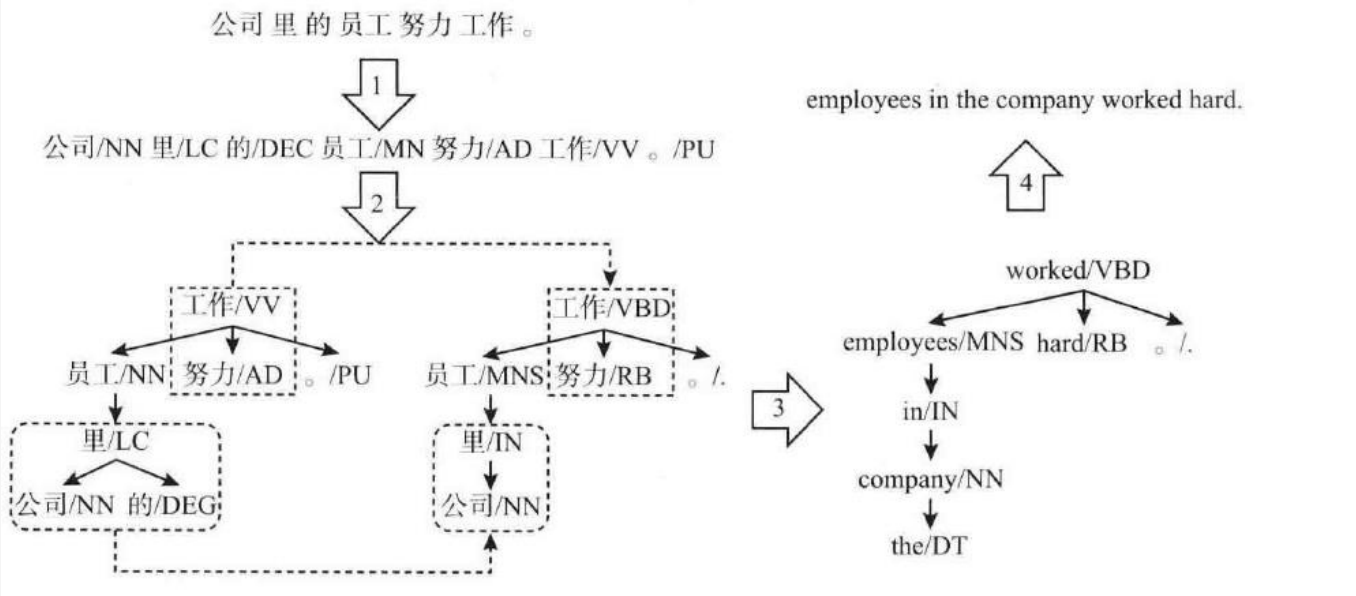


引文链接：https://zhuanlan.zhihu.com/p/60718978

基于规则的机器翻译

机器翻译基于规则的翻译，翻译知识来自人类专家。找人类语言学家来写规则，这一个词翻译成另外一个词。这个成分翻译成另外一个成分，在句子中的出现在什么位置，都用规则表示出来。这种方法的优点是直接用语言学专家知识，准确率非常高。

基于规则的机器翻译通常分为源语言句子分析、源语言句子转换和目标语言句子生成三个阶段，如图所示：



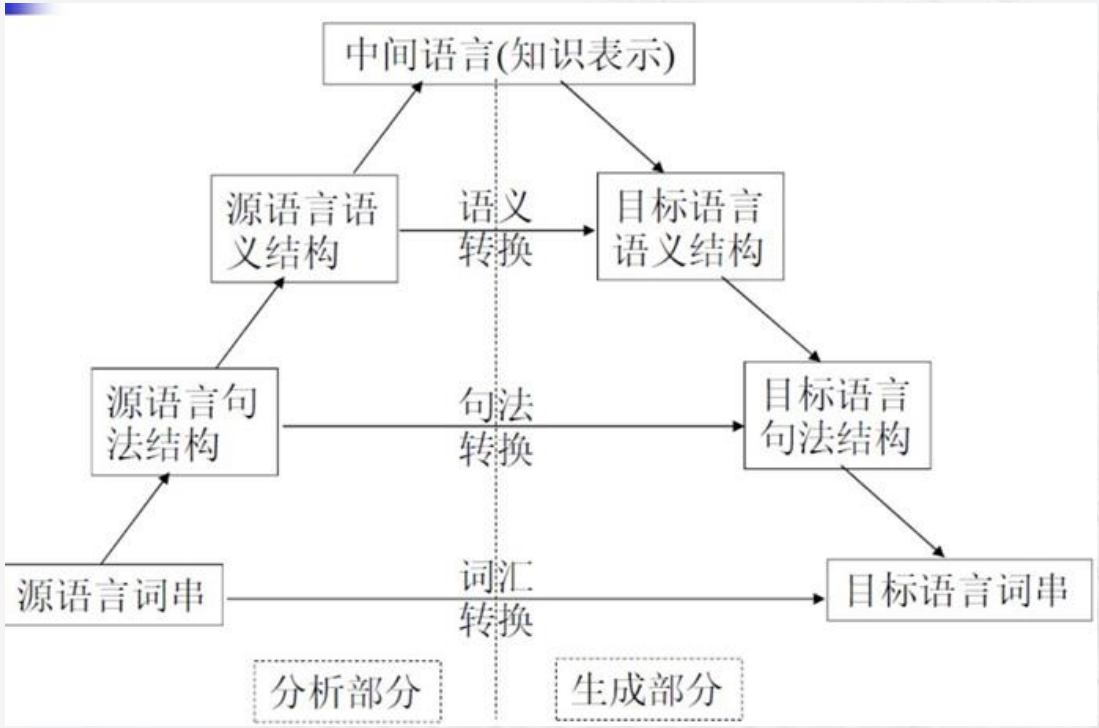
给定输入的源语言句子经过词法和句法分析得到句法树，然后通过转换规则将源语言句子句法树进行转换，调整语序、插入词或者删除词并将句法树中的源语言词用对应的目标语言词替换，生成目标语言的句法树。最后基于目标语言的句法树遍历叶子节点，得到目标语言句子。

基于规则的机器翻译

机器翻译是符号主义的典型方法。

- 建模时：
  - 建立“中间语言”符号体系；
  - 获取大量翻译实例集合；
  - 总结翻译规则。

- 翻译时：
  - 将源语言分析为中间语言；
  - 根据规则将中间语言转为目标语言。



基于规则的机器翻译

实例

① 源语言的词法分析

她把一束花放在桌上。⟹ She put a bunch of flowers on the table.

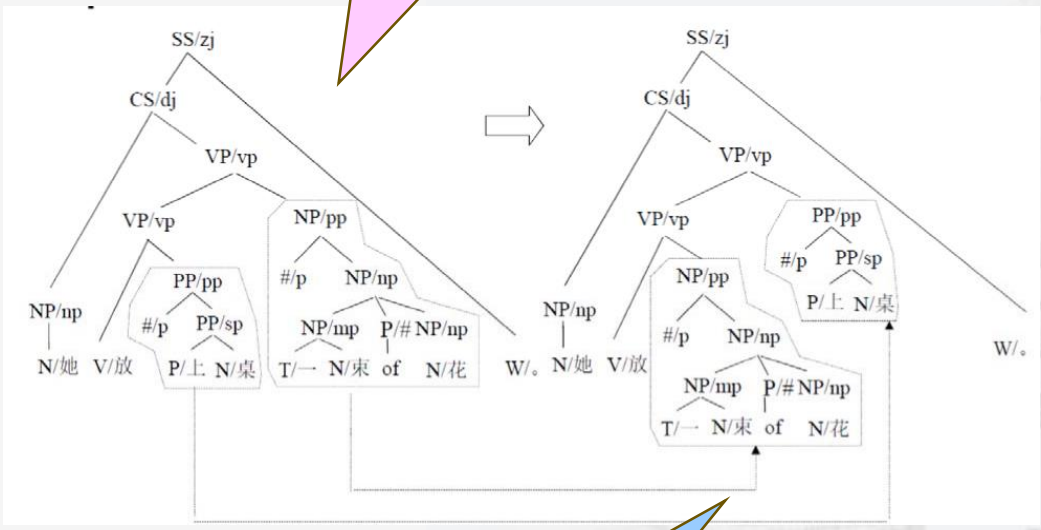
Segmenter/POS Tagger

她/r 把/p-q-v-n 一/m-d 束/q 花/n-v-a 放/v 在/p-d-v 桌/n 上/f-v 。/w

filter

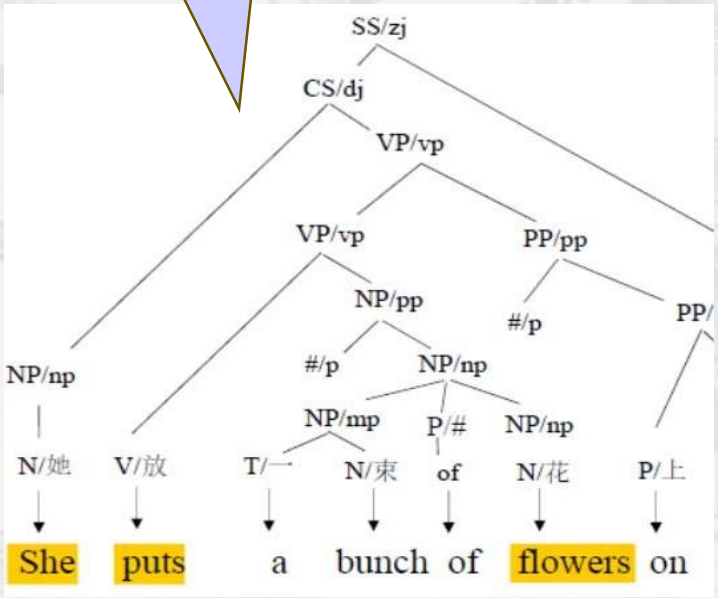
她/r 把/p 一/m-d 束/q 花/n 放/v 在/p-v 桌/n 上/f-v 。/w

② 源语言的句法分析



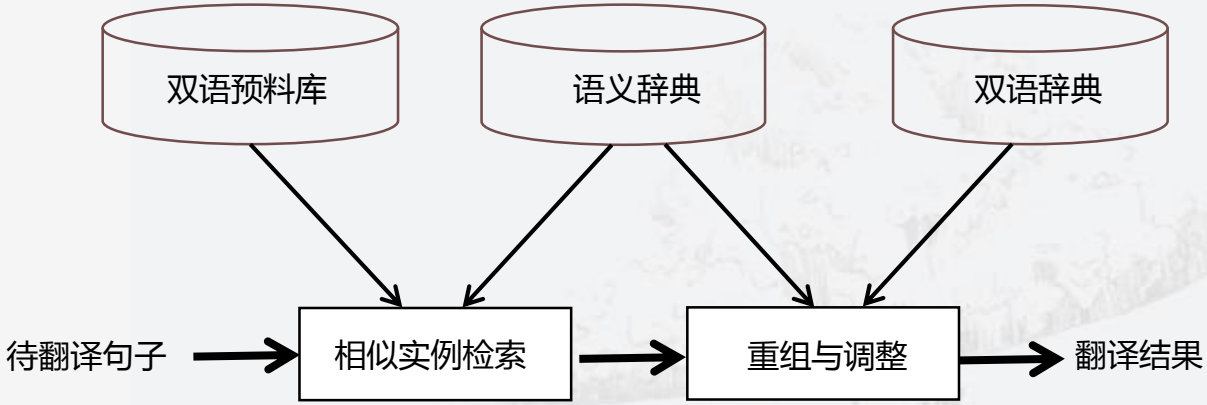
③ 匹配句法转换规则

④ 生成目标语言



基于实例的机器翻译

人们收集一些双语和单语的数据，并基于这些数据**抽取翻译模板以及翻译词典**。在翻译时，计算机对输入句子进行翻译模板的匹配，并基于匹配成功的模板片段和词典里的翻译知识来生成翻译词典，如图所示：



- 优点：
  - 高效性：通过复用已有的翻译实例，可以快速完成翻译任务。
  - 灵活性：能够适应不同的语言结构和表达方式。
  - 准确性：通过变形操作，能够提高翻译的准确性。

- 缺点：
  - 依赖高质量的双语语料库：是基础但其构建和维护成本较高。
  - 适用范围有限：对于新出现的词汇和表达方式，基于实例的机器翻译可能无法有效处理。

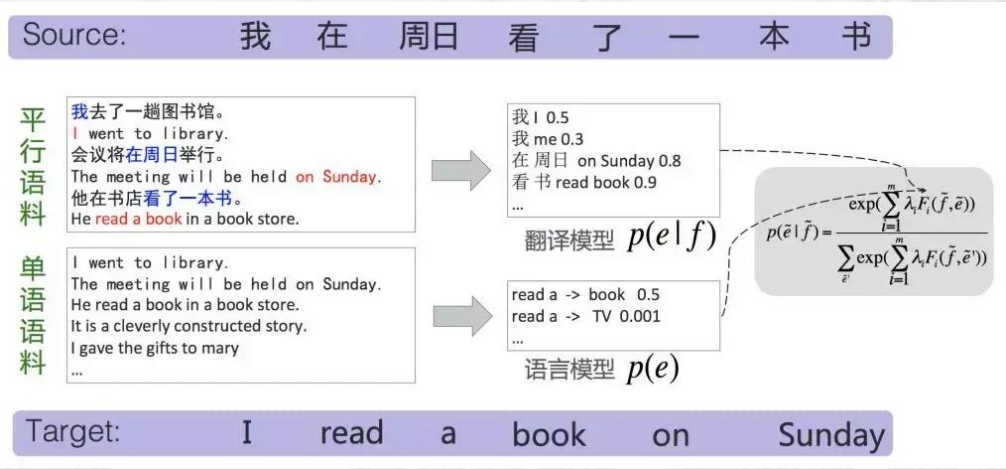
基于大规模语料的统计方法

基于统计的机器翻译，根据给定源语言句子，对目标语言句子的**条件概率进行建模**，通常拆分为**语言模型和翻译模型**，翻译模型刻画目标语言句子跟源语言句子在意义上的一致性， 而语言模型刻画目标语言句子的流畅程度。它的成本是非常低的，因为这个方法是语言无关的。一旦这个模型建立起来以后，对所有的语言都可以适用。

翻译知识主要来自两类训练数据：平行语料，一句中文一句英文，并且这句中文和英文，是互为对应关系的，也叫双语语料；单语语料，比如说只有英文我们叫单语语料。

翻译模型从平行语料中能能学到类似于词典这样的一个表，一般称为『短语表』。比如说『在周日』可以翻译成『on Sunday』。后面还有一个概率，衡量两个词或者短语对应的可能性。这样，『短语表』就建立起两种语言之间的一种桥梁关系。

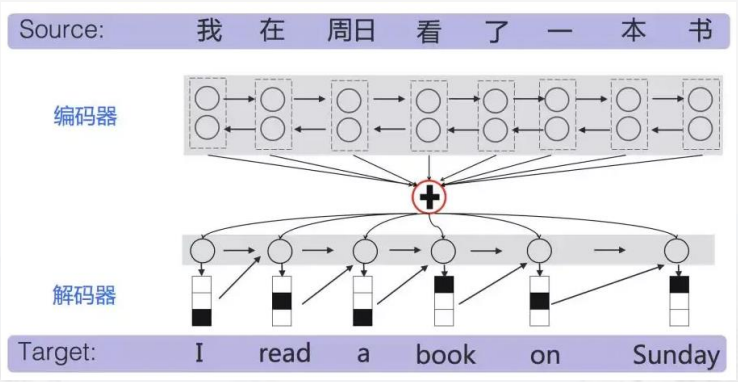
我们用单语语料来训练语言模型，语言模型就是衡量一个句子在目标语言中是不是地道，是不是流利。比如这里说『read a book』，这个表述是没有问题的，『read a 』后面跟一个『book 』这个词的概率可能是0.5，那么如果说『read a TV』呢？可能性就很低。因为这不符合目标语言的语法。



基于随着互联网的快速发展，大规模的双语和单语语料的获取成为可能，基于大规模语料的统计方法成为机器翻译的主流。统计机器翻译通常使用某种解码算法生成翻译候选，然后用语言模型和翻译模型对翻译候选进行打分和排序，最后选择最好的翻译候选作为译文输出。解码算法通常有束解码、CKY 解码等。

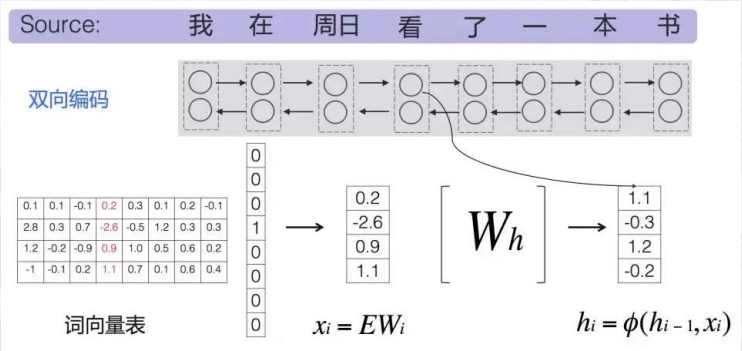
神经网络机器翻译

相比统计机器翻译而言，神经网络翻译从模型上来说相对简单，它包含两个部分，一个是**编码器**，一个是**解码器**。编码器是把源语言经过一系列的神经网络的变换之后，表示成一个高维的向量。解码器负责把这个高维向量再重新解码（翻译）成目标语言。



编码器进行了一个双向的编码，就是把词用词向量来表示，词向量表是通过神经网络训练出来的。源语言句子中的词，可以用一个one hot的向量表示。所谓one hot就是，比如例中句子有8个词。哪个词出现了，就把这个词标为1，其他的词标为0。比如第4个词“看”这个词是1，那么其他的都是0。

这两个矩阵相乘，相当于一个查表的操作，就把其中这个词向量表的一列取出来了，那么这一列的向量就代表了这个词。



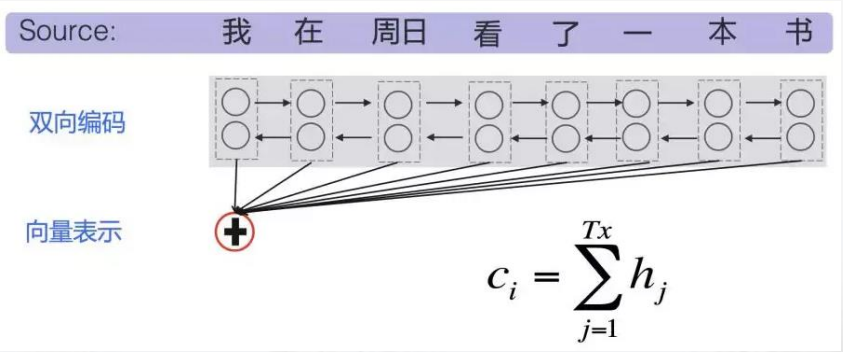
神经网络里面所有的词都会用向量来表示。得到词的向量表示后，再经过一个循环神经网络的变换，得到另外一个向量，称为Hidden State（隐状态）。

为什么做了一个双向的编码？是为了充分利用上下文信息。比如说，如果只是从左往右编码，“我在周日看”，看的是什么呢？“看”后面的你不知道，因为你只得到了“看”前面的信息。那么怎么知道后面的信息呢，这时候我们就想那能不能从后面到前面再进行一个编码，那就是“书本一了看”，从后面往前的编码，这时候“看”呢既有前面的信息，也有后面的信息。所以它有了一个上下文的信息，可以进一步提高译文质量。

神经网络机器翻译

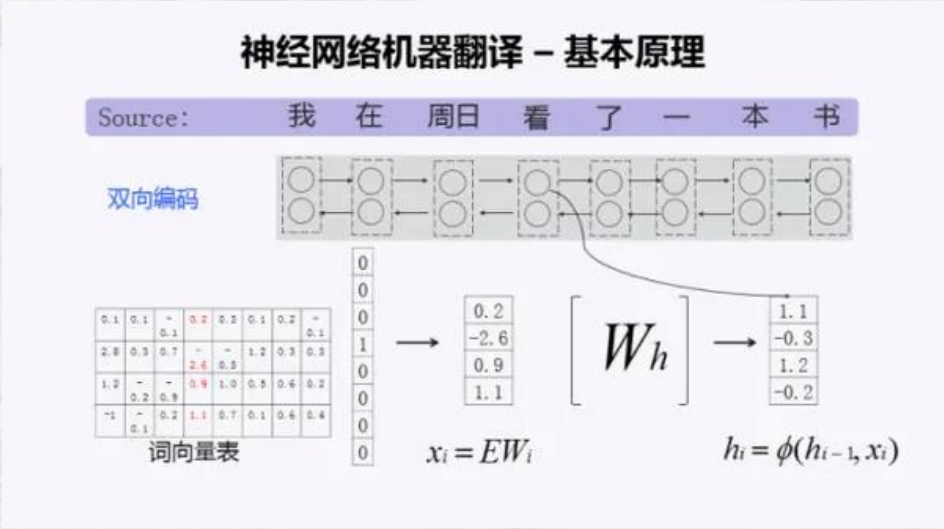
每个词经过一系列变换，映射为一个向量表示。如果将双向编码的向量结合起来呢？现在一般采用将两个向量进行拼接。比如两个256维的向量，拼接完成后得到一个512维的向量，用来表示一个词。

编码器编码完成后，需要把这个源语言的句子压缩到一个向量里去。最简单的方式是把所有的向量加起来。



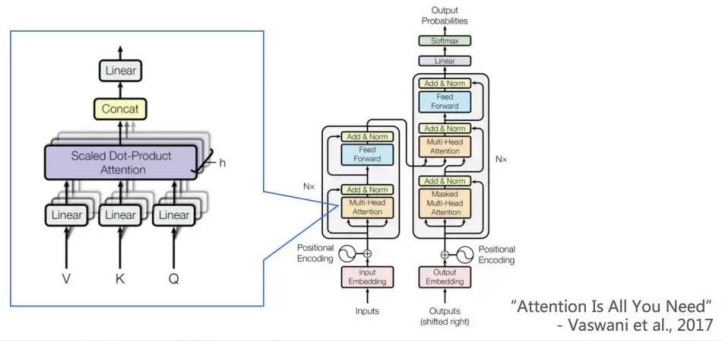
每一个词都是被作为相同的权重去对待的，那显然是不合理的，这时候就提出了一个**注意力机制，叫Attention**。这里用不同深度颜色的线去表示Attention的能量强弱，用以衡量产生目标词时，它所对应的源语言词的贡献大小。所以呢h前面又加一个α，α就表示它的一个权重。

有了句子的向量表示后，就掌握了整个源语言句子的所有的信息。解码器就开始从左到右一个词一个词的产生目标句子。在产生某个词的时候，考虑了历史状态。第一个词产生以后，再产生第二个词，直到产生句子结束符EOS(End of Sentence)，这个句子就生成完毕了。

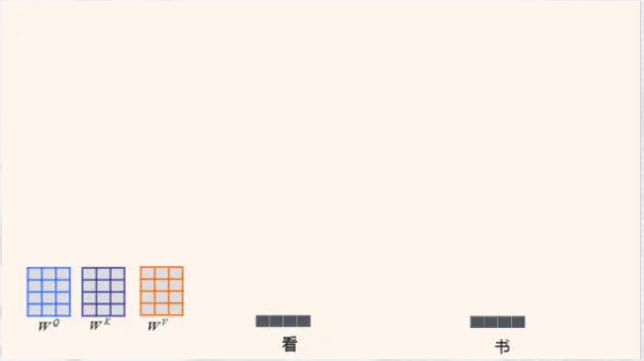


神经网络机器翻译

TRANSFORMER基本上取得了目前来说神经网络机器翻译最好的效果。TRANSFORMER的改进在哪里，来源一篇论文——“Attention Is All You Need”，论文所提出的方法可以只用注意力机制就把翻译搞定了。它其实也有一个编码器和一个解码器，这个是架构是没有变的。其中编码器和解码器都有多层。下面我们通过一个具体例子，来简单解释一下其原理。

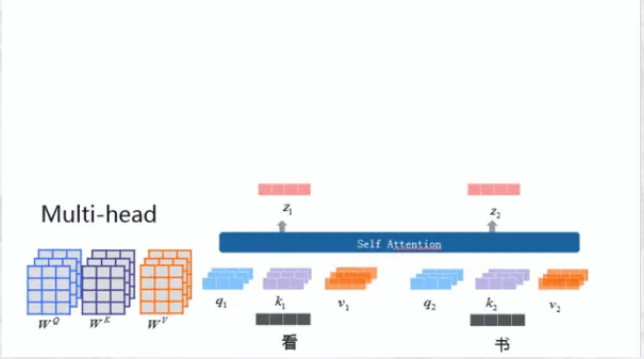


我们这个句子就包含两个词『看书』。论文中，把每一个词都用三个向量表示，一个叫Query（Q），一个叫Key（K），另外一个是Value（V）。那怎么得到一个词的Query、Key和Value呢？左边有三个矩阵，WQ、WK和WV，只要跟每一词向量相乘，就能够把这个词转换成三个向量表示。目标是把『看』这样一个词，通过一系列的网络变换，抽象到高维的向量表示。



通过Q和K进行点积，并通过softmax得到每个词的一个attention权重，在句子内部做了一个attention，称作Self Attention。Self Attention可以刻画句子内部各成分之间的联系，比如说“看”跟“书”之间就建立了联系。这样，每个词的向量表示（Z）就包含了句子里其他词的关联信息。

论文认为只有这一个QKV不太够，需要从多个角度去刻画，提出了“Multi-head”。在里面论文里面定义了8组QKV的矩阵（这个数值可以自定义）。在通过一系列变换，最终得到了每个词的向量表示。这只是encoder一层。那么这一层的输出做为下一层的输入，再来一轮这样的表示，就是Encoder-2，那么再来一轮就是第三层，如此一直到第N层。Decoder也是类似，不再解释。



机器翻译的评价方式

译价机器翻译的译文质量主要有两种方式。

第一种，人工评价。一说人工评价，大家第一时间就会想到『信、达、雅』，这是当年严复老先生提出来。我们用『信』来衡量忠实度，语言是为了交流的，『信』衡量译文是不是忠实地反映了原文所要表达的意思。『达』可以理解为流利度，就像刚才语言模型那样衡量的，译文是不是在目标语言中是一个流畅、地道的表达。至于『雅』，相对比较难衡量，这是仁者见仁、智者见智的。目前来说，机器翻译水平还远没有达到可以用『雅』来衡量的状态。

第二种，自动评价。现在一般采用的方法是基于n-gram（n元语法）的评价方法（用BLEU值）。BLEU是在多个句子构成的集合（测试集）上计算出来的。有了这个测试集以后，需要有参考答案（reference）。所谓参考答案就是人类专家给出的译文。这个过程很像考试，通过比较参考答案和系统译文的匹配程度，来给机器翻译系统打分。

Source: 我在周日读了一本书。
Reference: I read a book on Sunday.

System1: On Sunday, I read book.
System2: I read a book on Sunday.

BiLingual Evaluation Understudy

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad \text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

(Papineni et al., 2002)

显见，system2的得分会更高，因为它的译文跟reference是完全匹配的，system1匹配了一些片段，但是不连续。在计算BLEU得分的时候，连续匹配的词越多，得分越高。当然，BLEU值也有比较明显的缺点，BLEU分数受测试领域、reference多样性等多种因素的影响。Reference越多样化，匹配上的可能性就会越大。一般来说，抛开具体的设置，单说一个分数不具有参考性。

机器翻译的挑战

- **漏译。**句子有些词没有被翻译出来，甚至有时候一个长句子有逗号分隔，有几个子句没有翻译出来。漏译与词语的熵成正相关关系，词的熵越大，漏译的可能性越大，它所对应的目标语言词越多，概率越分散（熵越大），越有可能被漏译。
- **数据稀疏。**相比于统计机器翻译，这个问题对神经网络翻译而言更严重。针对数据稀疏问题，提出了一个多任务学习的多语言翻译模型，源语言共享编码器，在解码端，不同的语言，使用不同的解码器，这样在源语言端就会共享编码器的信息，从而缓解数据稀疏问题。。
- **引入知识。**如何将更多丰富的知识引入翻译模型是机器翻译长期面临的挑战。但目前引入知识还是比较表层的。知识的引入，还需要更多更深入的工作。比如『中巴』在没有给上下文的时候，是无法判断『巴』是哪个国家的简称。
- **可解释性。**可解释性（Interpretability）指的是人类理解和解释AI模型决策过程的能力。虽然人们可以设计和调整网络结构，去优化系统，提高质量，但是对于该方法还缺乏深入的理解。
- **语篇翻译。**这是机器翻译长期以来面临的挑战，大部分的翻译系统现在所使用的翻译方法都是基于句子，以句子作为单位，一个句子一个句子的进行翻译，单看句子翻译还可以接受，但是连起来看就觉得生硬不连贯。

西安交通大学

XI'AN JIAOTONG UNIVERSITY

6.3 自然语言人机交互

对话系统

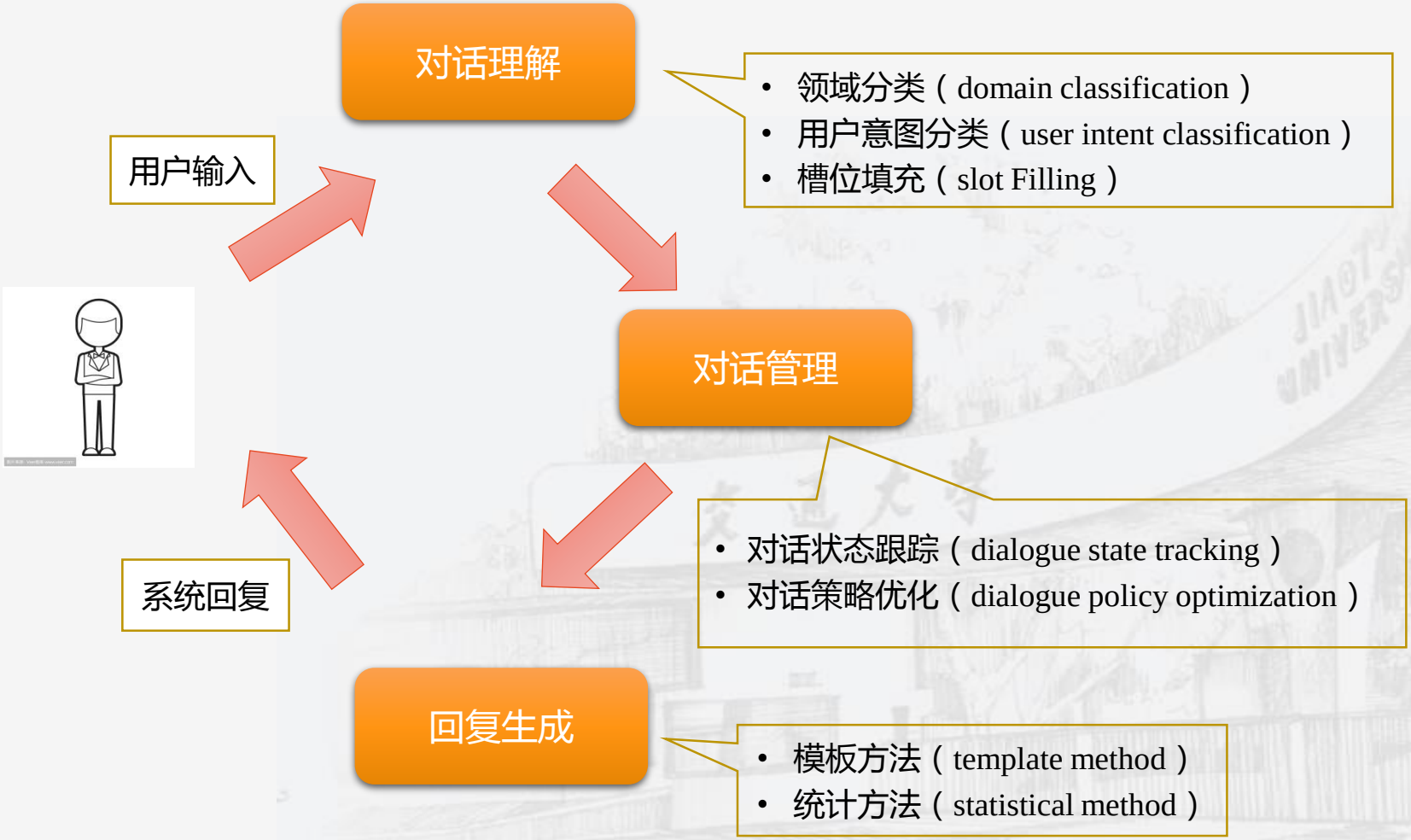
对话系统是指以完成特定任务为主要目的的人机交互系统。早期的对话系统大多以完成单一任务为主。近年来，面对多任务的对话系统不断涌现并且越来越贴近人们的日常生活。大多数对话系统由三个模块构成——**对话理解、对话管理和回复生成**。

- **对话理解。**首先，对话理解模块根据历史对话记录对用户当前输入的对话内容进行语义分析，识别出对话任务的领域（如航空领域）和用户意图（如机票预定），并抽取出完成当前任务所必需的若干必要信息（如起飞时间、起飞城市、到达城市、航空公司等）。
- **对话管理。**然后，对话系统根据用户当前输人的自然语言理解结果，对整个对话状态进行更新，并根据更新后的对话状态决定接下来系统需要采取的行动指令。
- **回复生成。**最后，回复生成模块基于对话管理输出的系统行动指令生成自然语言回复，并返回给用户。上述过程迭代进行，直到对话系统获取足够的信息并完成任务为止。

语音识别(ASR)和文本生成语音(TTS)也是对话系统的重要组成部分，前者负责将用户输人的语音信号转换成自然语言文本，后者负责将对话系统生成的自然语言回复转成语音。由于本章侧重自然语言部分，因此跳过对ASR和TTS的介绍。

对话系统

对话系统的流程图如下：



对话系统

对话理解

对话理解模块负责对用户输入的对话内容进行包括**领域分类、用户意图分类和槽位填充**在内的语义分析任务。

- **领域分类（domain classification）**：根据用户对话内容确定任务所属的领域。例如，常见的任务领域包括餐饮、航空和天气等。
- **用户意图分类（user intent classification）**：根据领域分类的结果进一步确定用户的具体意图，不同的用户意图对应不同的具体任务。例如，餐饮领域中常见的用户意图包括餐厅推荐、餐厅预定和餐厅比较等。

领域分类和用户意图分类同属分类任务，因此二者可以采用同一套方法完成。早期的分类方法主要基于统计学习模型，如最大熵和支持向量机等。近年来，基于深度学习的分类模型被广泛用于领域分类和用户意图识别任务。如基于深度信念网络(deep belief net)的分类方法、基于深度凸网络(deep convex network)的分类方法、基于循环神经网络和卷积神经网络的分类方法等。这类方法无须人工指定特征，能够针对分类任务直接进行端到端的模型优化，并且在大多数分类任务上已经取得了最好的效果。

- **槽位填充（slot filling）**：针对某个具体任务，从用户对话中抽取出完成该任务所需的槽位信息。例如，餐厅预定任务所需的槽位包括就餐时间、就餐地点、餐厅名称和就餐人数等。

槽位填充属于序列标注任务，每个任务对应的槽位信息由一系列键 -值对构成。每个键(key)对应一个具体槽位，例如餐厅预定任务中的就餐时间、就餐地点、餐厅名称和就餐人数等;每个值(vale)对应当前槽位对应的具体赋值。

对话系统

对话管理

对话管理模块主要由对话状态跟踪和对话策略优化两部分组成。前者负责在每轮对话结束时对整个对话状态进行动态更新，后者负责根据更新后的对话状态决定接下来系统将采取的行动。对话管理方法可以分为三类：

1）基于有限状态机的方法

将对话过程看成是一个有限状态转移图，通过使用槽位填充输出的键-值对更新对话状态（包括对某个槽位加入对应的值，以及更新或删除某个槽位对应的历史值），并根据当前状态转移图的状态决定接下来将要采取的行动。该类方法优点是可以通过对目标任务的理解制定明确清晰的状态转移图，并采用基于规则的方法控制对话过程；缺点是真实对话中往往会出现诸如反复询问或插人题外话的异常情况，该方法缺乏对此类异常的有效对应机制。

2）基于部分可观测马尔科夫过程的方法

该类方法基于真实对话数据，将语音识别和自然语言理解模块的不确定性引入模型。该方法将对话过程看作是一个马尔可夫决策过程，并用转移概率P(s_t |$s\_(t-1)$,$a\_(t-1)$)来表示从对话状态$s_(t-1)$到对话状态s_t之间的转移。这里的每个对话状态s，对应一个变量，该变量无法直接观察到。POMDP将自然语言理解模块的输出o_t看作是带有噪音的、基于用户输入的观察值，这个观察值的概率为P(s_t |o_t)。

3）基于深度学习的方法

将神经网络用于对话状态跟踪任务。在该类方法中，对话状态跟踪模块负责对整个对话历史和系统目前对话状态进行编码，并基于该编对对整个对话状态进行更新；对话策略优化模块采用增强学习技术决定接下来系统需要采取的行动指令。这类方法通过最大化未来回报(future reward) 的方式进行上述两个模型的参数优化，并根据训练好的模型生成最优的行动指令。

对话系统

回复生成

回复生成模块负责根据对话管理模块输出的系统行动指令，生成对应的自然语言回复并返回给用户。典型的回复生成方法包括**基于模板**的方法和**基于统计**的方法两类。

- **其于模板的方法**使用规则模板完成从系统行动指令到自然语言回复的转化，规则模板通常由人工总结获得。该类方法能够生成高质量回复，但模板扩展性和句子多样性明显不足。
- **基于统计的方法**完成从系统行动指令到自然语言回复的转化。基于规划的（plan-based）方法通过句子规划（sentence planning）和表层实现（surface realization）两步完成转化任务。句子规划负责将系统行动指令转化为某种预定义的中间结构，表层实现负责将该中间结构进步转化为自然语言回复并输出给用户。这类方法缺点在于句子规划阶段依然需要使用预先设计好的规则。

基于语料（corpus-based）的方法有效地缓解了上述问题。例如，基于语言模型的方法从系统行动指令出发，基于语言模型直接生成自然语言回复句子，而不再经过任何中间状态。这类方法的优点在于尽量少地避免了对人工规则的过度依赖。不过，传统语言模型对长距离依存现象的处理存在天然的不足。基于深度学习的方法使用基于循环神经网络的序列生成模型替换了语言模型，不仅能够有效解决长距离依存问题，而且借助深度学习能够自行选择特征的机制，能够采用端到端的方式在任务数据上直接优化。

聊天机器人

聊天机器人是一种人工智能交互系统，其工作方式是通过语音或文字实现人机在任意开放话题上的交流。目前，人们建立聊天机器人的目的在于模拟人类的对话行为，从而检测人工智能程序是否能够理解人类语言并且和人类进行长时间的自然交流，使用户沉浸于对话环境之中。

聊天机器人技术大致可分为三类：基于规则的聊天机器人、基于检索的聊天机器人和基于生成的聊天机器人。

1）基于规则的聊天机器人

设计者会预先定义好一系列的规则，给定对话输入，首先规则系统对输入进行自然语言解析，在解析过程中抽出预定义的关键词等信息；之后根据所抽取的关键信息，通过定义好的模板进行回复。如果输入不在规则体系之内，则用万能回复进行回复用户。

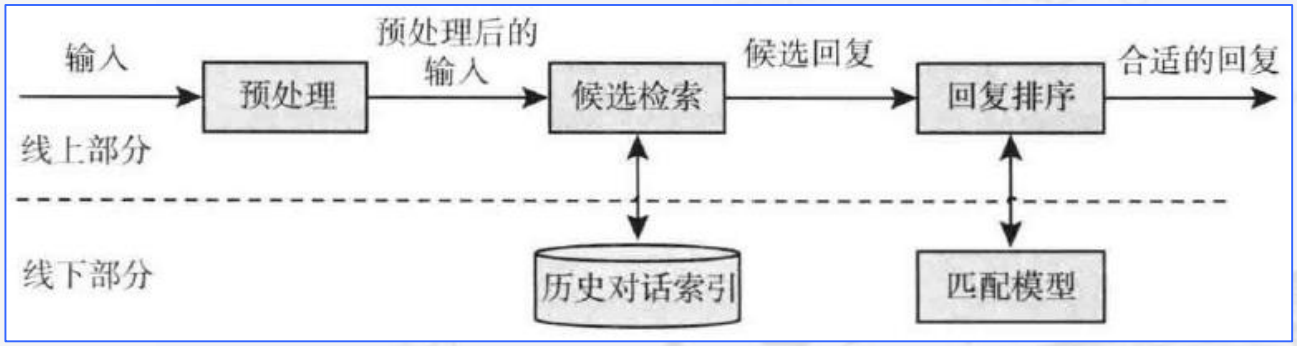
优点：回复可控，每条回复均由设计者撰写，并且回复触发的逻辑也被精心设计。

缺点：难以覆盖所有开放领域的聊天话题，很多话题没有合适的回复，系统的可扩展性较弱。

聊天机器人

2）基于检索的聊天机器人

利用成熟的搜索引擎技术和人类对话语料构建的聊天机器人系统。检索式聊天机器人分为线上和线下两部分。线下部分由索引、匹配模型以及排序模型三个模块组成，这三个模块分别为线上产生候选回复、信息—回复对的特征描述以及回复候选的排序。索引中收集了大量来自社交网络上人与人的交流数据，组织成“一问一答”结构。



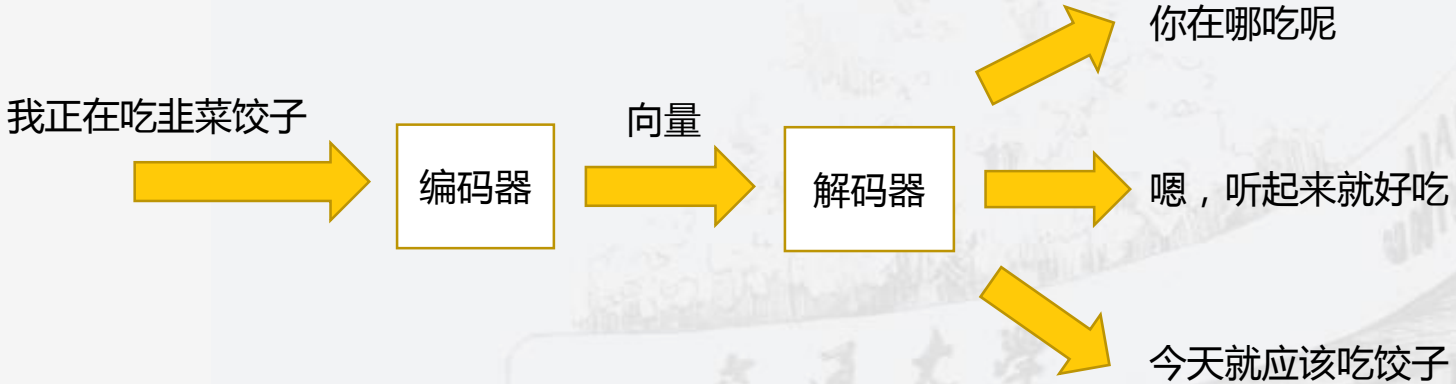
匹配模型是检索式聊天机器人的关键，其作用是实现对用户信息和回复候选的语义理解，对二者语义上构成回复关系的可能性进行打分。这些打分在线上构成了每个信息-回复对的特征，而这些特征最终由一个排序模型进行整合，产生最终的候选排序。

检索式聊天机器人的本质是对已有的人类回复进行筛选重用来回复新的信息，所以回复的好坏很大程度上依赖于索引的质量和是否能够检索到合适的候选，并且检索式聊天机器人没有显式地将人类常识建模到系统之中。因此，如何检索到能和上下文逻辑一致的回复候选以及如何精准分析上下文，均是当前检索式聊天机器人所面临的挑战。

聊天机器人

3）基于生成的聊天机器人

利用自然语言生成技术对给定对话上下文直接生成一句完整的话语进行回复。此类算法可以基于已有模型产生训练集中没有出现过的回复。生成式聊天机器人目前普遍基于神经网络的“序列到序列”模型实现。



与机器翻译类似，生成式聊天机器人中的“序列到序列”仍然需要海量聊天数据进行训练才能产生良好的回复。然而尽管如此，生成模型的很多回复缺乏信息或者过于普适(如“是啊”“我也觉得”此类)。产生这种回复的主要原因是人类对话十分复杂，没有明显的对应关系，而且很多回复需要额外的人类知识才能生成。

在机器翻译中，一个源语言一般只有有限的几种翻译；而在对话中，特别是聊天机器人的开放域对话中，一条输入信息可以有上千种合适的回复。这种过于倾斜的“一对多”的对应关系使得在机器翻译中表现良好的编码-解码模型只能捕捉到对话中少数高频模板，从而产生普适回复。普适回复一方面会降低回复的相关性，另一方面也会使人和机器的聊天很难进行下去。另外，如何自动衡量一个生成模型的好坏仍然是一个值得探索的问题。

西安交通大学
XI'AN JIAOTONG UNIVERSITY

6.4 智能问答

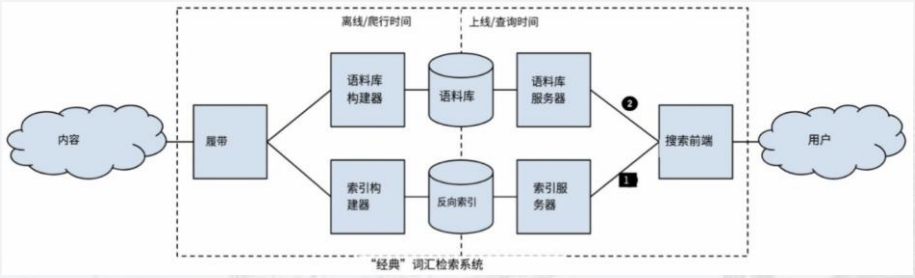
智能问答

智能问答旨在**模拟人类对话者回答用户问题的能力**，其核心功能是**理解用户**提出的问题，并从海量信息中**筛选**出最相关的答案，甚至在一些情况下**生成**合适的答案。

智能问答系统可以分为两类：

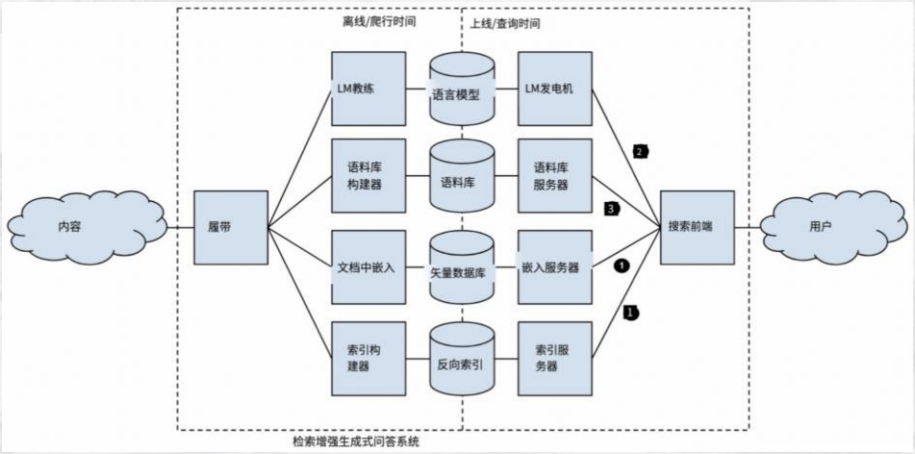
① 封闭域问答系统（Closed-domain QA）：该系统仅在特定领域或主题下回答问题，它在某一特定领域拥有深入的知识，并专门针对这些领域的知识库进行问答。

右图展示了一种闭环生成式问答系统的架构，其中语言模型是唯一的知识存储库。这个模型由一个离线组件进行更新和训练，然后由一个在线组件用于实际的问题生成和回答



② 开放域问答系统（Open-domain QA）：该系统可以回答来自任何领域的问题，依赖于广泛的知识库或数据库进行检索和回答。开放域系统通常依赖更为复杂的技术来处理和筛选信息。

右图是这种系统的高级示意图，使用了语义索引作为检索组件。这样做的优点是可共同优化检索、答案提取或生成，还可以通过增加一个词汇检索组件来增强语义检索，从而获得混合检索器。



引文链接：https://blog.csdn.net/lt5227/article/details/140352998

智能问答

工作步骤

1）问题理解与解析：问答系统首先需要理解用户的提问。这个过程包括：

- **语法分析：**通过自然语言处理技术，将问题进行语法分析，识别出问题中的关键字、短语及其关系。
- **意图识别：**通过深度学习模型，识别用户提问的意图。例如，用户可能想要知道某个概念的定义。
- **实体识别：**识别出问题中的实体（如人名、地名、日期、数字等），以便进一步理解问题的语义。

2）信息检索与候选答案生成：理解问题后，问答系统进入信息检索阶段。具体步骤如下：：

- **知识库检索：**如果是封闭域问答系统，系统会从事先构建的知识库中检索相关的信息。如果是开放域问答系统，通常会依赖于更广泛的互联网资源（如百度百科、新闻网站等）。
- **文档检索：**通过关键词匹配、语义匹配等方式，检索到可能包含答案的文档或段落。检索方法通常基于TF-IDF、BM25等传统的检索模型，或者基于深度学习的模型（如BERT等）进行语义相似度匹配。
- **候选答案生成：**从检索到的文档中，提取出相关的句子或段落，作为候选答案。

3）答案排序与生成：候选答案生成后，需对答案进行排序并选择最优答案。该过程包括：

- **答案评分：**使用机器学习模型对候选答案进行评分，确定哪些答案更符合问题的需求。评分机制可以基于语义相似度、答案的完整性和相关性等因素。
- **答案生成：**在某些情况下，尤其是开放域问答系统中，模型不仅需要从现有文档中提取信息，还需要生成合适的回答。基于生成的模型（如GPT系列、T5等）可以将问题和检索到的信息结合起来，生成符合语义的自然语言答案。

4）答案呈现：

最后，问答系统将最优答案呈现给用户。用户通常希望得到简洁明了的答案，而不是冗长的解释。因此，如何优化答案的表达方式，也是智能问答系统面临的一大挑战。

西安交通大学
XI'AN JIAOTONG UNIVERSITY

谢谢大家