阅读笔记

不定期记录近期阅读的文章，记下对文章的理解和问题，思考下一步阅读的方向。

目录

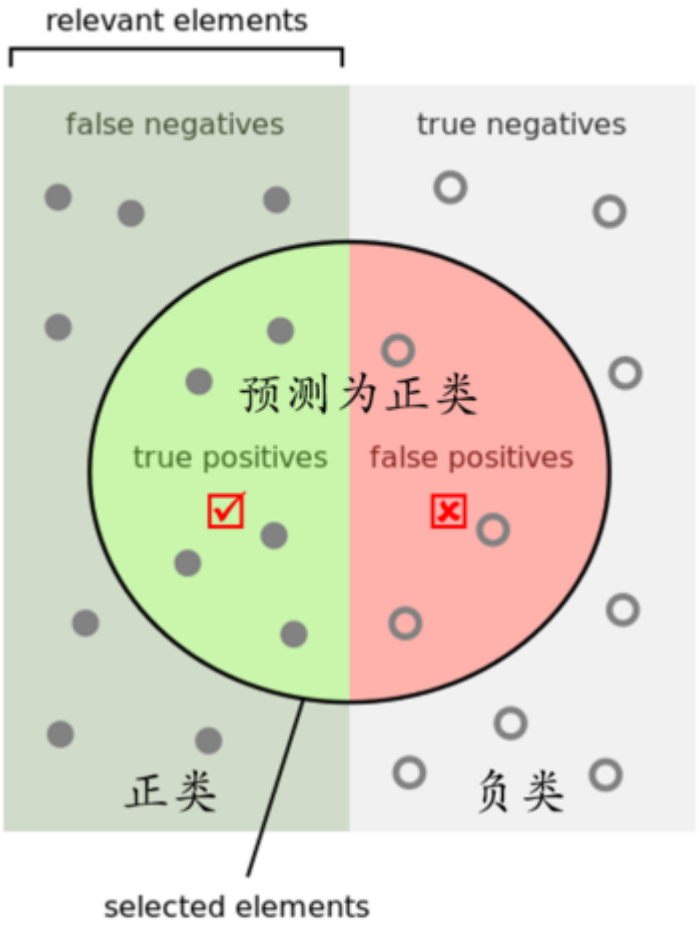
2016-10-20 序列标注

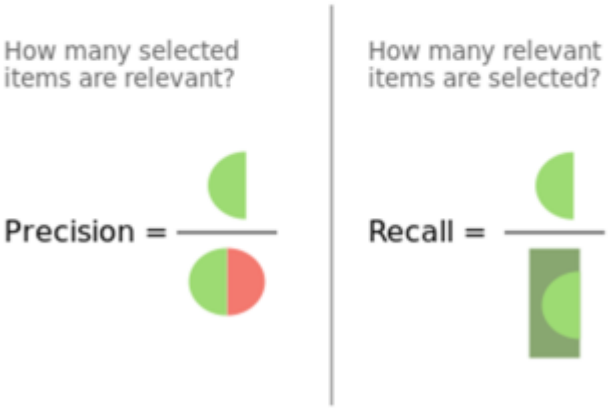
A. 近期阅读文章

1. 机器学习性能评估指标
2. 误差模型：过拟合，交叉验证,偏差-方差权衡
3. 使用RNN解决NLP中序列标注问题的通用优化思路
4. CRF和LSTM 模型在序列标注上的优劣？ - 回答作者: 谢志宁
5. 序列标注模型
6. 标注偏置问题(Label Bias Problem)和HMM、MEMM、CRF模型比较
7. Comparisons of sequence labeling algorithms and extensions
8. 聊天机器人中对话模板的高效匹配方法

B. 个人理解

- 精确率(precision)和准确率(accuracy)是不一样的，在正负样本不平衡的情况下，准确率这个评价指标有很大的缺陷。比如在互联网广告里面，点击的数量是很少的，一般只有千分之几，如果用acc，即使全部预测成负类（不点击）acc也有99%以上，没有意义。
- 精确率是针对预测结果而言的，它表示的是预测为正的样本中有多少是对的。那么预测为正就有两种可能了，一种就是把正类预测为正类(TP)，另一种就是把负类预测为正类(FP)。
  召回率是针对原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。那也有两种可能，一种是把原来的正类预测成正类(TP)，另一种就是把原来的正类预测为负类(FN)。

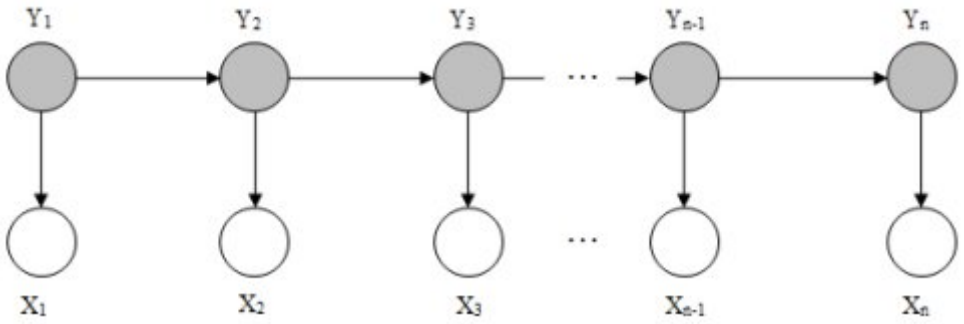




F1值是精确率和召回率的调和均值。

- 在序列标注问题上，RNN等模型在序列建模上很强大，它们能够capture长远的上下文信息，此外还具备神经网络拟合非线性的能力，这些都是CRF无法超越的地方，对于t时刻来说，输出层yt受到隐层ht（包含上下文信息）和输入层xt（当前的输入）的影响，但是yt和其他时刻的yt'是相互独立的，对当前t时刻来说，我们希望找到一个概率最大的yt，但其他时刻的yt'对当前yt没有影响，如果yt之间存在较强的依赖关系的话（例如，形容词后面一般接名词，存在一定的约束），RNN无法对这些约束进行建模，模型的性能将受到限制。
  而CRF模型更多考虑的是整个句子的局部特征的线性加权组合（通过特征模版去扫描整个句子）。关键的一点是，CRF的模型为p(y | x, w)，注意这里y和x都是序列，它有点像list wise，优化的是一个序列y = (y1, y2, …, yn)，而不是某个时刻的y_t，即找到一个概率最高的序列y = (y1, y2, …, yn)使得p(y1, y2, …, yn| x, w)最高，它计算的是一种联合概率，优化的是整个序列（最终目标），而不是将每个时刻的最优拼接起来，在这一点上CRF要优于RNN。

- 做为一种概率图模型，CRF在理论上更完美一些，一步一步都有比较坚实的理论基础。RNN理论上是能拟合任意函数的，对问题的假设明显放宽了很多，不过深度学习类模型的理论原理和可解释性一般。
- RNN可以当做对序列的一种『中间状态』的建模，建模结果还可以当做特征，与CRF结合使用。
- 关于产生式模型与判别式模型的概念，二者在分类器中经常被提及：假定输入X，类别标签Y：产生式模型估计联合概率P(x,y)，判别式模型估计条件概率P(y|x)。产生式模型可以根据贝叶斯公式得到判别式模型，但反过来不行。
- 现有的序列标注模型主要有HMM，MEMM 以及 CRF。
    1. 隐马尔科夫模型（HMM）：HMM是一种产生式模型，采用一个联合概率将观测序列和状态序列结合在一起，通过训练参数来最大化拟合训练语料的联合概率。

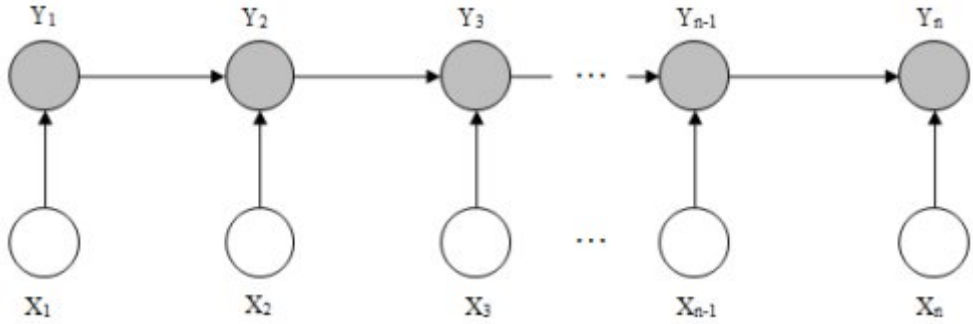


HMM是一个双重随机过程：马尔可夫链，用来描述状态的转移；随机过程，用来描述状态值和观察值之间的统计关系。

缺点：

1. 为对联合概率分布进行建模，HMM引入两条独立性假设：第一，马尔科夫链在任意时刻的状态yi仅依赖于前一个状态yi-1；第二，任意时刻的观测xi只依赖于该时刻马尔科夫链的状态yi。这就导致HMM只能局限于部分上下文特征，无法充分利用更多有效的特征。
2. 为定义观察值和状态值的联合概率，产生式模型必须列出所有可能的观察序列，在实际操作中很难实现。

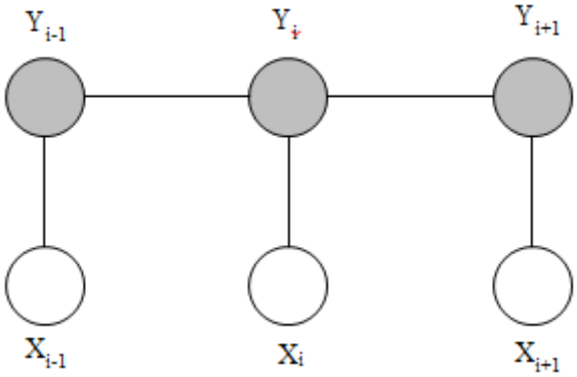
2. 最大熵隐马尔科夫模型（MEMM）：MEMM是一种判别式模型，不需要HMM那样严格的独立性假设。MEMM是基于概率有限状态模型这样一个概念，该模型将观察序列看作是条件事件，而不是由状态生成的。它结合了MEM和HMM的优点，允许状态转移可以基于输入序列中的非独立性特征，使得MEMM在处理自然语言处理的任务时，性能优于HMM。MEMM是通过求局部最优的条件概率来获得最终的条件概率。



观察序列是作为条件，而不是生成的，因此图的分布指的是t时刻状态Yi所表示的随机变量的联合分布。

缺点：仅对局部求解条件概率，取其概率最大的标注作为最终的输出标注，导致标注偏置问题的产生，即凡是训练语料中未出现的情况全都忽略掉。文章6中详细介绍标注偏置问题(Label Bias Problem)。

3. 条件随机场模型（CRF）：条件随机场也是一种判别式模型，是指在给定输入节点条件下计算输出节点的条件概率，其核心思想是利用无向图理论使序列标注的结果达到在整个序列上全局最优。



CRF无需引入独立性假设，能够充分利用上下文信息特征；计算全局最优输出节点的条件概率，克服了最大熵马尔可夫模型存在的标记偏置问题。
缺点：训练代价大、特征函数复杂度高。

- HMM模型是对转移概率和表现概率直接建模，统计共现概率。MEMM模型是对转移概率和表现概率建立联合概率，统计的是条件概率，容易陷入局部最优。CRF模型统计了全局概率，考虑了数据在全局的分布，而不是仅仅在局部归一化，解决了MEMM中的标记偏置问题。
- 文章7中比较了一些模型和算法在词性标注和OCR任务上的性能，包括HMM、CRF、AP、Structured SVM、M3N、SEARN算法，文中融合几个模型自创了SLE算法，进一步提升了相同条件下的模型性能。作者将所有模型参数调到最优，做两组试验，一组词性标注，一组手写字母OCR，在错误率和训练时间上看Structured SVM是性能最好的。
- 文章8中介绍了一种通过构建多模式匹配算法高效匹配对话模板的方法。在domain中缺少语料的情况下，高效的模板匹配算法在可行性上应该更高。

C. 存留问题

D. 接下来调研方向

返回目录

2016-09-19 深度学习

A. 近期阅读文章

1. 大牛的《深度学习》笔记，Deep Learning速成教程
2. Infographic: Rise of the Chatbots
3. 真正从零开始，TensorFlow详细安装入门图文教程！
4. 如何评价苹果的 AirPods 无线耳机？ - 回答作者: Tristan Zhang

B. 个人理解

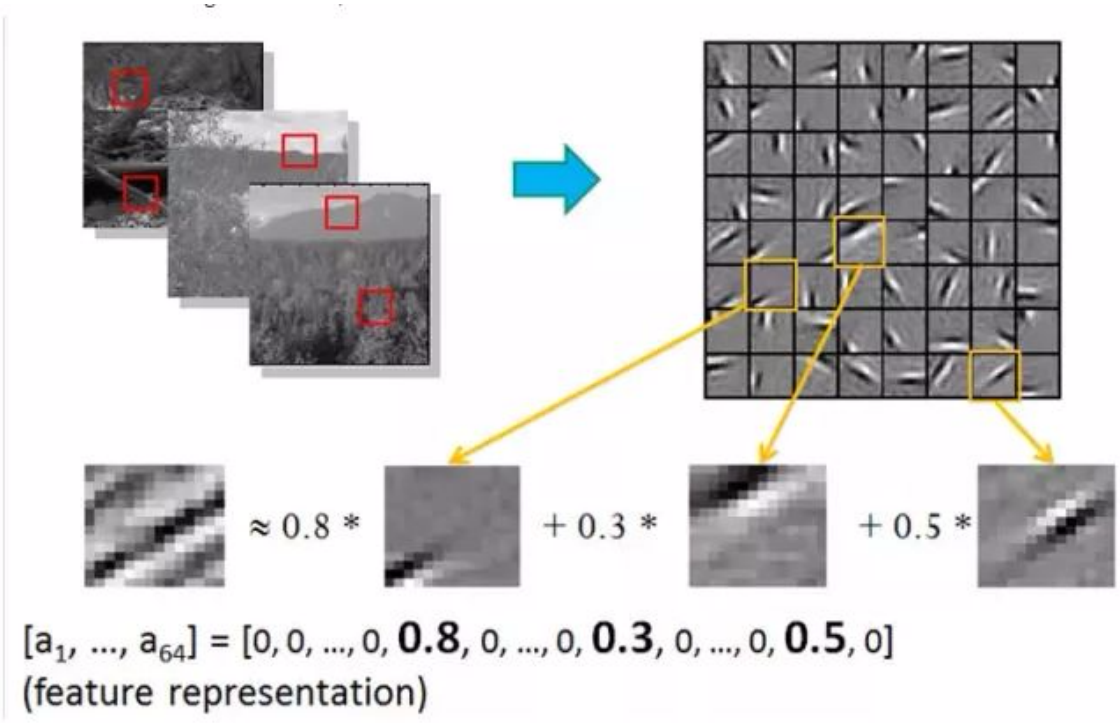
- 特征是机器学习系统的原材料，特征表示的粒度需要大到一定程度，学习算法才能发挥作用。
  1995年前后，Bruno Olshausen和David Field收集了很多黑白风景照片，从这些照片中，提取出400个小碎片，每个碎片的尺寸均为16x16像素，标记为S[i], i=0,..399，再从这些黑白风

景照片中，随机提取另一个碎片，尺寸也是16x16像素，标记为T。他们提出的问题是，如何从这400个碎片中，选取一组碎片，S[k] ,通过叠加的办法，合成出一个新的碎片与随机选择的目标碎片T尽可能相似，同时，S[k]的数量尽可能少。即：Sum_k(a[k] * S[k]) –> T, 其中a[k]是在叠加碎片S[k]时的权重系数。他们为解决这个问题发明稀疏编码（Sparse Coding）算法。

稀疏编码是一个重复迭代的过程，每次迭代分两步：

1. 选择一组S[k]，然后调整a[k]，使得Sum_k(a[k] * S[k])最接近T。
2. 固定住a[k]，在400个碎片中，选择其它更合适的碎片S’[k]，替代原先的S[k]，使得Sum_k(a[k] * S’[k])最接近T。

- 经过几次迭代后，最佳的S[k]组合被遴选出来了。令人惊奇的是，被选中的S[k]，基本上都是照片上不同物体的边缘线，这些线段形状相似，区别在于方向。



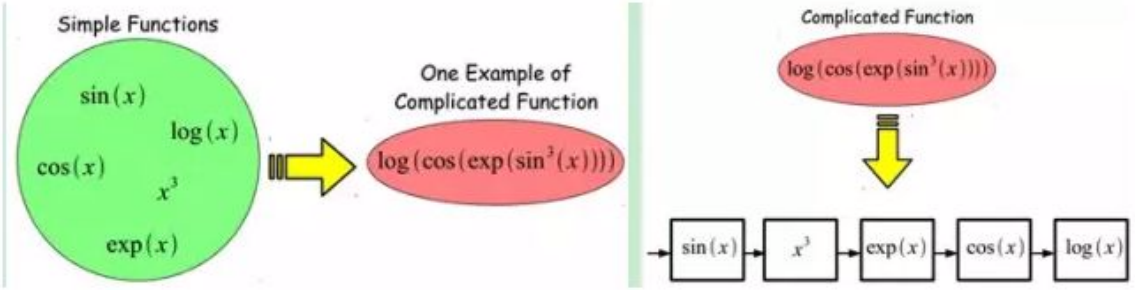
也就是说，复杂图形往往由一些基本结构组成，比如上图：一个图可以通过用64种正交的edges（可以理解成正交的基本结构）来线性表示。比如样例的x可以用1-64个edges中的三个按照0.8,0.3,0.5的权重调和而成。而其他基本edge没有贡献，因此均为0。同样的，声音也存在这种规律。

- 小块的图形可以由基本edge构成，更结构化，更复杂的，具有概念性的图形需要更高层次的特征表示，底层表达作为高层表达的基（basis），层次递进组合成更高层的表达。直观上说，就是找到make sense的小patch再将其进行combine，就得到了上一层的feature，递归地向上learning feature。
- 假设一个系统S，它有n层（S1,…Sn），它的输入是I，输出是O，形象地表示为：I =>S1=>S2=>…..=>Sn => O，如果输出O等于输入I，即输入I经过这个系统变化之后没有任何的信息损失，表示在任何一层Si，它都是原有信息（即输入I）的另外一种表示。在Deep Learning系统中通过调整每一层的参数，使得输出依旧等于输入，那么我们就可以自动地获取得到输入I的一系列层次特征，即S1，…, Sn。所以深度学习的思想就是堆叠多个层，上一层的输出作为下一层的输入，通过这种方式实现对输入信息进行分级表达。
- 20世纪80年代末期，用于人工神经网络的反向传播（Back Propagation, BP）算法可以让一个人工神经网络模型从大量训练样本中学习统计规律，从而对未知事件做预测，比起过去基于人工规则的系统在很多方面显出优越性。这个时候的人工神经网络，虽也被称作多层感知机（Multi-layer Perceptron），但实际是种只含有一层隐层节点的浅层模型。20世纪90年代，各种各样的浅层机器学习模型相继被提出，例如支撑向量机（SVM，Support Vector

Machines）、Boosting、最大熵方法（如LR，Logistic Regression）等。这些模型的结构基本上可以看成带有一层隐层节点（如SVM、Boosting），或没有隐层节点（如LR）。这些模型无论是在理论分析还是应用中都获得了巨大的成功，但是由于理论分析的难度大，训练方法又需要很多经验和技巧，这个时期浅层人工神经网络反而相对沉寂。深度学习在学术界和工业界的掀起浪潮是在2006年，加拿大多伦多大学教授、机器学习领域的泰斗Geoffrey Hinton和他的学生RuslanSalakhutdinov在《科学》上发表了一篇文章，有两个主要观点：

1. 多隐层的人工神经网络具有优异的特征学习能力，学习得到的特征对数据有更本质的刻画，从而有利于可视化或分类；
2. 深度神经网络在训练上的难度，可以通过“逐层初始化”（layer-wise pre-training）来有效克服，在这篇文章中，逐层初始化是通过无监督学习实现的。

- 深度学习可通过学习一种深层非线性网络结构，实现复杂函数逼近，表征输入数据分布式表示，并展现了强大的从少数样本集中学习数据集本质特征的能力。（多层的好处是可以用较少的参数表示复杂的函数）



深度学习的实质，是通过构建具有很多隐层的机器学习模型和海量的训练数据，来学习更有用的特征，从而最终提升分类或预测的准确性。因此，“深度模型”是手段，“特征学习”是目的。区别于传统的浅层学习，深度学习的不同在于：

1. 强调了模型结构的深度，通常有5层、6层，甚至10多层的隐层节点；
2. 明确突出了特征学习的重要性，通过逐层特征变换，组合低层特征形成更加抽象的高层表示属性类别或特征，将样本在原空间的特征表示变换到一个新特征空间，从而使分类或预测更加容易。与人工规则构造特征的方法相比，利用大数据来学习特征，更能够刻画数据的丰富内在信息。

- BP算法作为传统训练多层网络的典型算法，采用迭代的算法来训练整个网络，随机设定初值，计算当前网络的输出，然后根据当前输出和label之间的差去改变前面各层的参数，直到收敛（整体是一个梯度下降法），深度结构（涉及多个非线性处理单元层）非凸目标代价函数中普遍存在的局部最小是训练困难的主要来源。BP算法存在的问题：
    1. 梯度越来越稀疏：从顶层越往下，误差校正信号越来越小；
    2. 收敛到局部最小值：尤其是从远离最优区域开始的时候（随机值初始化会导致这种情况的发生）；
    3. 一般只能用有标签的数据来训练：但大部分的数据是没标签的，而大脑可以从没有标签的的数据中学习；

- 为了克服神经网络训练中的问题，DL采用了与神经网络不同的训练机制。
  如果对所有层同时训练，时间复杂度会太高；如果每次训练一层，偏差就会逐层传递。这会面临跟监督学习中相反的问题，会严重欠拟合（因为深度网络的神经元和参数太多了）。
  2006年，hinton提出了在非监督数据上建立多层神经网络的一个有效方法，分为两步，一是每次训练一层网络，二是调优，使原始表示x向上生成的高级表示r和该高级表示r向下生成的x’尽可能一致。方法是：
    1. 逐层构建单层神经元，这样每次都是训练一个单层网络。
    2. 当所有层训练完后，Hinton使用wake-sleep算法进行调优。

- 将除最顶层的其它层间的权重变为双向的，这样最顶层仍然是一个单层神经网络，而其它层则变为了图模型。向上的权重用于“认知”，向下的权重用于“生成”。然后使用Wake-Sleep算法调整所有的权重。让认知和生成达成一致，也就是保证生成的最顶层表示能够尽可能正确的复原底层的结点。比如顶层的一个结点表示人脸，那么所有人脸的图像应该激活这个结点，并且这个结果向下生成的图像应该能够表现为一个大概的人脸图像。Wake-Sleep算法分为醒（wake）和睡（sleep）两个部分:
    1. wake阶段：认知过程，通过外界的特征和向上的权重（认知权重）产生每一层的抽象表示（结点状态），并且使用梯度下降修改层间的下行权重（生成权重）。也就是“如果现实跟我想象的不一样，改变我的权重使得我想象的东西就是这样的”。
    2. sleep阶段：生成过程，通过顶层表示（醒时学得的概念）和向下权重，生成底层的状态，同时修改层间向上的权重。也就是“如果梦中的景象不是我脑中的相应概念，改变我的认知权重使得这种景象在我看来就是这个概念”。

- deep learning具体训练过程:
    1. 使用自下上升非监督学习（从底层开始，一层一层的往顶层训练）：
       采用无标定数据（有标定数据也可）分层训练各层参数，这一步可以看作是一个无监督训练过程，是和传统神经网络区别最大的部分（这个过程可以看作是feature learning过程）
       具体的，先用无标定数据训练第一层，训练时先学习第一层的参数（这一层可以看作是得到一个使得输出和输入差别最小的三层神经网络的隐层），由于模型capacity的限制以及稀疏性约束，使得得到的模型能够学习到数据本身的结构，从而得到比输入更具有表示能力的特征；在学习得到第n-1层后，将n-1层的输出作为第n层的输入，训练第n层，由此分别得到各层的参数；
    2. 自顶向下的监督学习（通过带标签的数据去训练，误差自顶向下传输，对网络进行微调）：
       基于第一步得到的各层参数进一步fine-tune整个多层模型的参数，这一步是一个有监督训练过程；第一步类似神经网络的随机初始化初值过程，由于DL的第一步不是随机初始化，而是通过学习输入数据的结构得到的，因而这个初值更接近全局最优，从而能够取得更好的效果；所以deep learning效果好很大程度上归功于第一步的feature learning过程。

- 语音助手放到手机里，打开手机才能唤醒确实有点蠢，成功的科技应该让用户感觉不到生硬，而是自然而然的与之交互，苹果尝试取消耳机接口，以后可能用蓝牙耳机唤醒可能是一个不错的方式，但是感觉Amazon Echo那种就放在家里，可能已经忘记它的存在了，但是想到有什么任务要做时，不用做任何前置准备一句话就唤醒的语音助手，随时问想问的问题，安排需要的任务，这更接近一个科幻中描述的人工智能。

C. 存留问题

D. 接下来调研方向

- DST部分具体工作过程

返回目录

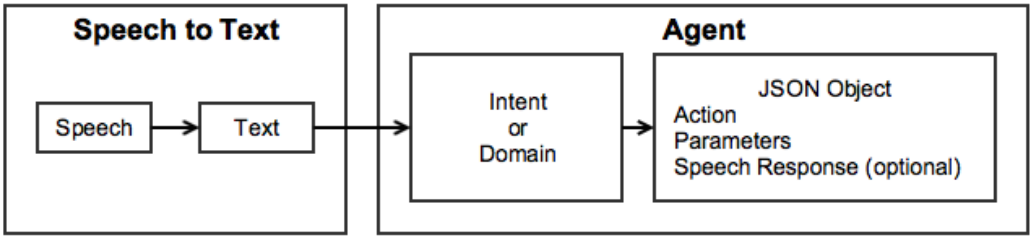
2016-09-07 神经网络、机器表示

A. 近期阅读文章

1. 口语对话系统中对话管理方法研究综述
2. 从api-ai工作原理来看构建简单场景chatbot的一般方法
3. 深度｜人工神经网络深入分析
4. 教机器学习表示
5. Efficient Estimation of Word Representations in Vector Space

B. 个人理解

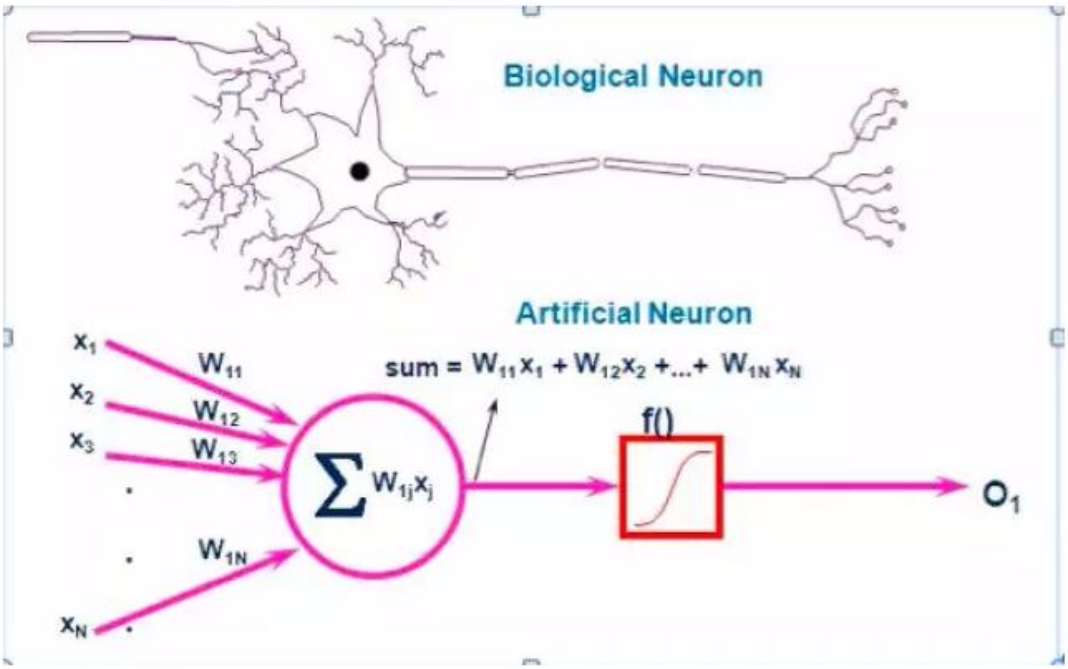
- api.ai（最近被Google收购）工作原理：



  Chatbot的NLP部分包括NLU、DST和NLG，其中NLU是DST的核心和基础，而DST是chatbot的核心。api.ai提供了一个简单场景Chatbot的解决方案，重点解决了NLU问题，从user inputs中识别出user intent和对应的action，然后从user inputs中抽取出预先设定好的entity value，作为action的parameter传给developer后台执行，返回结果用来填充预先定义好的templates完成NLG任务。Intents、Actions、Entities需要developer从特定任务中归纳，分别编写Intents、Entities的examples用来训练chatbot准确地识别user intents和entity parameters，api.ai提供了训练器，也可以developer自己写算法实现。
- api.ai是面向developer的Chatbot开发工具，提供web界面，使用简单，提高了Chatbot开发的效率，在NLG和Context分析部分还有待提高，但在总体上降低了开发chatbot的门槛，是个很有意义和前景的服务。
- 1943年，基于生物神经网络莫克罗-彼特氏神经模型（McCulloch - Pitts'neuronmodel）诞生，由心理学家Warren McCulloch和数学家Walter Pitts合作提出，基本思想是抽象和简化生物神经元的特征性成分，不需要捕捉神经元的所有属性和行为，但要足以捕获它执行计算的方式。

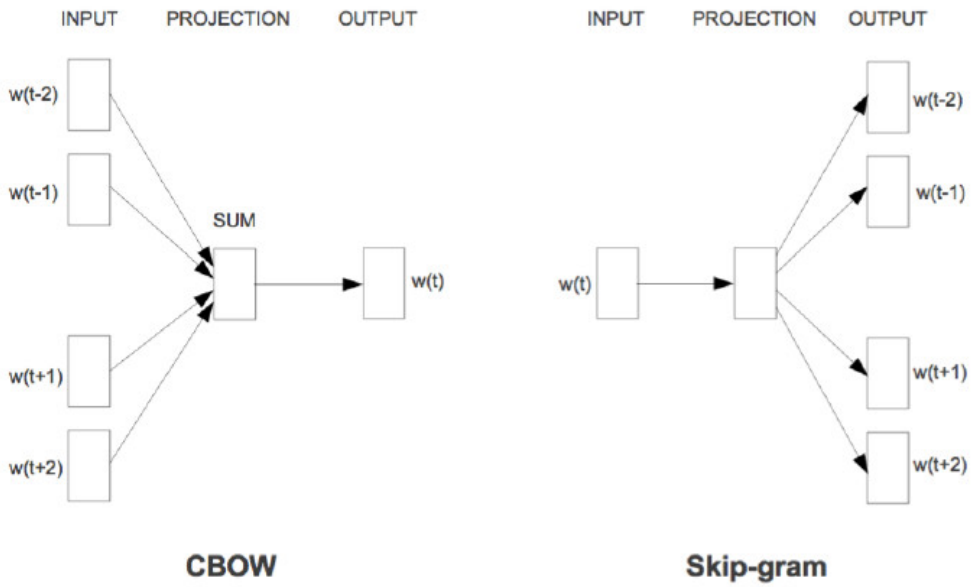
生物神经元与 MP 模型

生物神经元	神经元	输入信号	权值	输出	总和	膜电位	阈值
MP 模型	j	χ_i	ω_{ij}	o_j	$\sum$	$\sum_{i=1}^{n}\omega_{ij}\chi_i(t)$	T_j



时间离散，时刻t（t=0，1，2，……）得到兴奋型输入xi，如果膜电位等于或大于阈值以及抑制型输入为0，则在时刻t + 1，神经元细胞输出为1否则为0。

- 罗森布拉特（Frank Rosenblatt）于1957年提出了”感知器” (Perceptron)，这是第一个用算法来精确定义的神经网络。感知器由两层神经元组成，输入层接收外界信号，输出层是McCulloch-Pitts神经元，即阈值逻辑单元。1969 年，Marvin Minsky和Seymour Papert出版了新书《感知器：计算几何简介》中论证了感知器模型的两个关键问题：其一，单层的神经网络无法解决不可线性分割的问题，典型例子如同或（XNOR，两个输入如果相同，输出为1；两个输入如果是不同，输出为0。），需要多层神经网络学习数据间的复杂关系。其二，受硬件水平的限制，当时的电脑完全没有能力完成神经网络模型所需要的超大的计算量。
- 模型的对比在数据集的不同、数据规模的不同、模型的不同、模型超参数的不同都会导致不同的结果和结论，data driven的模型效果的好坏与数据有着直接的关系，在工程应用上，还是需要在自己的数据上做对比实验，而不是用所谓的“理论”来剖析哪种模型更好。
- 文章[5]中介绍的CBOW模型用Message来预测word，将每个词赋以一个n维向量初值，以每个词向量求和来表示Message，拿来预测目标词，不断地训练得到最终的词向量。SKIP-GRAM模型与CBOW类似，只是用word来预测context。



文章中主要强调了两个模型把词用50-100维的相对稠密的向量表示，降低了过去方法的计算复杂度，提高训练效率，并且可以计算词之间的相似关系。

C. 存留问题

D. 接下来调研方向

- 深度学习系统实现，TensorFlow

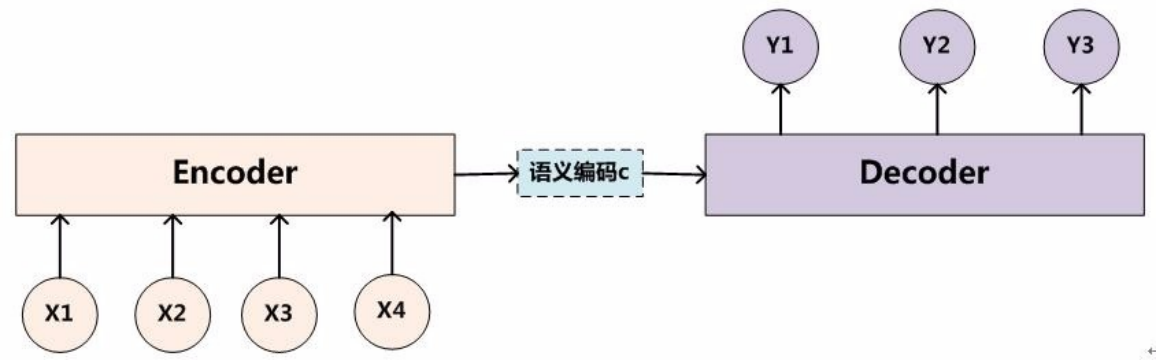
返回目录

2016-08-19 seq2seq模型

A. 近期阅读文章

1. 自然语言处理中的Attention Model：是什么及为什么
2. 使用深度学习打造智能聊天机器人
3. PaperWeekly 第二期
4. Deep Learning in NLP（一）词向量和语言模型
5. 词向量（Distributed Representation）工作原理是什么？ - 回答作者: 皮果提
6. A Neural Conversational Model
7. A Diversity-Promoting Objective Function for Neural Conversation Models
8. Thought Vectors, Deep Learning & the Future of AI
9. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing
10. Siri技术解析

B. 个人理解

- 

  Encoder-Decoder框架：适合处理由一个句子（或篇章）生成另外一个句子（或篇章）的通用处理模型。对于句子对<X,Y>，目标是给定输入句子X生成目标句子Y。X和Y可以是同一种语言，也可以是两种不同的语言。而X和Y分别由各自的单词序列构成：

  $$\mathbf{X} = \langle \mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_m \rangle$$

  $$\mathbf{Y} = \langle \mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_n \rangle$$

  Encoder顾名思义就是对输入句子X进行编码，将输入句子通过非线性变换转化为中间语义表示C：$\mathbf{C} = \mathcal{F}(\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_m)$

  对于解码器Decoder来说，其任务是根据句子X的中间语义表示C和之前已经生成的历史信息y1,y2….yi-1来生成i时刻要生成的单词yi：$\mathbf{y}_i = \mathcal{G}(\mathbf{C}, \mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_{i-1})$

- Encoder-Decoder是个非常通用的计算框架，至于Encoder和Decoder具体使用什么模型都是由研究者自己定的，常见的比如CNN/RNN/BiRNN/GRU/LSTM/Deep LSTM等，变化组合非常多，而很可能一种新的组合就是一个创新。文章[6]中Google采用的是2个LSTM。
- 生成模型问题与解决方案：
  1. 多轮对话问题：上下文聊天信息Context应该引入到Encoder中，因为这是除了当前输入Message外的额外信息，有助于Decoder生成更好的会话应答Response内容。深度学习解决这个问题时大致思路相同，都是在Encoder阶段把上下文信息Context及当前输入Message同时编码，但是对于RNN模型来说，如果输入的线型序列长度越长，模型效果越差，所以可以采用多层向前神经网络代替RNN模型，或者采用层级神经网络（Hierarchical Neural Network，HNN），上下文Context中每个句子首先用“句子RNN(Sentence RNN)”对每个单词进行编码形成每个句子的中间表示，而第二级的RNN则将第一级句子RNN的中间表示结果按照上下文中句子出现先后顺序序列进行编码，这级RNN模型被称作”上下文RNN（Context RNN）”。Encoder引入Context促进Decoder阶段可以参考上下文信息生成更符合情景的应答Response。文章[9]的Dynamic Memory Network（DMN）将上下文Context和背景知识存入Episodic Memory module，激活需要的vector放入Decoder中。
  2. 安全回答问题：不论用户说什么内容，聊天机器人都用少数非常safe，grammatical但没有实际意义的response，比如”I don’t know!”之类的。原因在于传统的seq2seq在decoding过程中都是以MLE(Maximum Likelihood Estimate)为目标函数，即生成最grammatical的话，而不是最有用的话，这些safe句子大量地出现在训练语料中，模型学习了之后，无可避免地总是生成这样的response，文章[7]借鉴了语音识别的一些经验，在decoding的时候用MMI（Maximum Mutual Information）作为目标函数，提高了response的diversity。
  3. 个性信息一致问题：Chatbot作为一个虚拟人物，相关的个性化信息比如年龄、性别、爱好、语言风格等应该维护回答的一致性。利用经典的Sequence-to-Sequence模型训练出的聊天助手往往很难保持这种一致性，因为Sequence-to-Sequence模型训练的都是单句Message对单句Response的映射关系，内在并没有统一维护聊天助手个性信息

的场所。解决这个问题的技术思路应该都是类似的，核心思想是把聊天助手的个性信息在Decoder阶段能够体现出来，以此达到维护个性一致性的目的。

- dialog state tracking（DST）的作用类似于一张user conversation logs状态表，记录着用户当前的状态，以订机票为例，这张表的key是预先设定好的slots，比如目的地、出发地、出发时间等等，与系统背后的业务数据表中的attributes相关联，不断地从user conversation中抽取相应的values来填充这个表格，或者将其定义为一个多分类任务，不断地从对话中判断这句话中包括哪些slots和values（这里的values是多个分类结果），当状态表中的信息存在空白时，bot会根据空白的slots来提问并获取values，直到获取到足够的slots，给出用户suggestion，或者进行相应的服务。
- 使用LSTM时自定义cell是一个很有启发性的思路，针对具体问题的特点，修改现有的cell结构，也许会起到非常关键的作用。
- 实际应用中常常遇到data规模太小的问题，DL难以发挥作用，但如果从大量相似的domain data中学习一些表示模型，然后迁移到待解决的问题上，混合大量的相似domain数据，会cover到更丰富的features，为DL提供了广阔的舞台。
- One-hot Representation：把每个词表示为一个很长的向量，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1，1的位置对应该词在词典中的位置，采用稀疏方式存储，会非常的简洁。但这种词表示有两个缺点：
    1. 容易受维数灾难的困扰，尤其是将其用于 Deep Learning 的一些算法时；
    2. 任意两个词之间是孤立的，不能很好地刻画词与词之间的相似性，即”词汇鸿沟”。

- DL中一般用到的词向量是Distributed representation，通常被称为”Word Representation”或”Word Embedding（词嵌入）”，通过训练将某种语言中的每一个词映射成一个固定长度的短向量（当然这里的“短”是相对于one-hot representation的”长”而言的，50维和100维比较常见），将所有这些向量放在一起形成一个词向量空间，而每一向量则为该空间中的一个点，在这个空间上引入”距离”，则可以根据词之间的距离来判断它们之间的（词法、语义上的）相似性了。向量的距离可以用最传统的欧氏距离来衡量，也可以用cos夹角来衡量。
- 利用神经网络算法生成词向量，词向量通常和语言模型捆绑在一起，即训练完后两者同时得到。文章[4]中详细介绍了从大量未标注的普通文本数据中无监督地学习出语言模型和词向量的几个最经典的工作。

C. 存留问题

- 词向量、语言模型训练和使用完整过程？

D. 接下来调研方向

- 阅读经典paper，完整推导具体过程，实践。

返回目录

2016-08-15 Chatbot产品比较

A. 近期阅读文章

1. Siri、IBM Watson、微软小冰、Messenger M、Google Allo……聊天机器人这么火，背后的技术你都知道吗？
2. 当我们谈论 Bot 的时候，我们在谈论什么
3. 微软陆奇：“小冰”做对的两件事
4. 谷歌Allo的“智商”哪儿来？人工智能技术大揭秘
5. Introducing DeepText: Facebook’s text understanding engine
6. pet,baby and bot

B. 个人理解

产品	定位	技术路线	备注
Siri	个人助理	模板+搜索	利用语音识别和自然语言处理对用户语音进行分析，提取出关键字填入固定任务模板，最后根据填充的模板通过Restful协议来调用服务接口完成任务或给用户反馈
Watson	领域知识问答	模板+NLU+DL+知识库	集成高级自然语言处理、信息检索、知识表示、自动推理、机器学习等开放式问答技术，基于为假设认知和大规模的证据搜集、分析、评价
微软小冰	娱乐聊天	NLU+DL+个性化+情感	在对话中感知上下文，支持多人对话，利用到了知识库、网络搜索和社交网络平台的搜索结果，通过整合和后处理得到更精确的回答
Messenger M	个人助理	模板+搜索+DL+RL	基于深度学习的文本解析引擎DeepText,不断地学习聊天内容中包括主题在内的语义知识
Allo	个人助理	模板+搜索+DL+RL+用户嵌入	采用一个两阶段深度学习模型，可通过用户输入建立用户画像，学习用户行为

- Chatbot需要增强的点：
    1. 建立用户与bot的关系：对用户进行建模，增强用户与机器之间的亲密感，有一定的辨别和判断。
    2. 多模态输入输出：表情、语气、动作，提升交互感。
    3. 记忆与推理：对用户的输入进行深层次的记忆与推理，有一定的归纳和整理。

- Chatbot领域火热的原因：
    1. 网络生态：消息服务类应用迅速壮大，几乎占领了用户所有的碎片时间，事实上成为了移动互联网时代的“浏览器”入口。用户注意力难以分散到其他类应用上，下载移动应

用所带来的流量红利慢慢消失。而对话系统开发成本低，可以依附于大平台，绕开应用下载量和活跃量的问题，成为消息服务之上最自然的应用，而且在移动时代成长起来的用户天然接受即时消息通讯的方式，进入门槛低、粘性高，或许可以取代移动应用构建自己的生态环境。

2. 技术变革：深度学习的飞速进展让多层神经网络得以在计算机视觉和语音识别领域取得突破性进展，让人工智能在原本人类擅长的领域表现的更为优异。深度学习的成果震撼到了人们，这项技术也将拓展到更多的领域研究中去，包括自然语言处理（NLP），对话系统可以承借这股东风一举突破现在面对的各种瓶颈，实现曾经遥不可及的梦想。
3. 自然语言作为人机交互界面：在理解文本语义的基础上自动进行下一步动作，很多繁琐、重复的文字类的人工劳动将被自动化的机器取代，释放出的市场潜力无疑是非常巨大的。

- Chatbot面临的困境：
    1. 在许多场景下 APP 的操作更加简单，Chatbot 并未体现出以自然语言作为交互界面的优势。
    2. 对于机器理解人们日常使用的自然语言这件事情，事实上我们与几年前相比并未取得明显的进步，目前的Chatbot还远远达不到人们对流畅对话的期待。

- 微软小冰做对的两件事：
    1. 情感计算框架：情感会影响决策，人工智能产品应当把EQ+IQ作为基本纬度。
    2. 通用对话系统：小冰是一套可以自我进化的人工智能开放域对话系统，很像是搜索引擎时代的长尾用户体验，不管用户抛出何种问题，小冰都能够回答，并把对话进行下去。小冰在对话中可以感知上下文，支持多人对话，利用到了知识库、网络搜索和社交网络平台来搜索答案，与传统的问题检索技术相比，用户得到的并不是知识库中固定的信息和一些网页，而是在搜索引擎的结果基础上通过整合和后处理得到更精确的回答。

- Google Allo采用一个非常创新的两阶段深度学习模型，可通过用户输入建立用户画像，学习用户行为。首先，一个递归神经网络一个字一个字地查看聊天语境，然后用长短时记忆（LSTM）的隐藏状态将其编码，这个编码作为一个矢量用来生成作为离散语义类别的回复。每一个语义类别都与它可能的回复库关联起来，然后使用第二个递归神经网络，加入了用户嵌入（user embbeding）特性，将softmax层转变为一个层级性的softmax层，使用beam搜索在大量可能回复库中选择出顶层得分最高的回复，最终将用户画像转换为一个隐藏的LSTM状态用来生成一个具体信息。
- Messenger M采用DeepText框架，充分利用了几个深度神经网络结构，包括卷积（Convolutional）和循环（Recurrent）神经网络，并能完成以单词、字符为基础的学习任务。通过可以理解用户输入内容，提取其中意图、情感和实体（人物、地点和事件），Messenger M可以代替用户完成任务，比如购物、送礼物、预约行程等，也可以在明确用户需求后从朋友的博文中分析出相关的内容进行推荐。
- bot应该像人和宠物一样具有自我学习能力，现在增强学习的思路和训练宠物很像，用一个reward作为牵引，带着bot学习programmer希望bot学习的action，未来bot或许可以像人一样尝试通过问问题来主动地学习，而不是仅仅回答问题。
- 迁移能力很重要，bot学习过类似的东西，就应该可以做类似的事情，而不是每次都需要重新从头开始学习，如何将已经学习到的知识迁移到新的领域是一个非常有意义的topic。

C. 存留问题

- 强化学习是怎么工作的？
- 深度学习应用到对话生成的具体过程？

D. 接下来调研方向

- 增强学习
- 深入理解神经网络在Chatbot上的应用

返回目录

2016-08-03 IBM Watson

A. 近期阅读文章

1. Google says machine learning is the future. So I tried it myself
2. 想了解机器学习？你需要知道的十个基础算法
3. Logistic Regression 模型简介
4. Introduction to “This is Watson”
5. 一张图带你看懂IBM Waston的工作原理
6. 5 tips for building the next great chatbot

B. 个人理解

- 机器学习正在重塑这个世界，它会向因特网技术一样，从无人问津到最终成为每个人都可以做一点的事情。过去我们尝试教会机器完成复杂的任务，而现在是构建出一个能够让机器自主学习如何完成任务的系统，行业先驱者利用这种方法重新完成了图片搜索、声音识别等一些目前使用量已经很大的服务，并从根本上让其重生，然后开源出自己的软件包让世界跟上他们的脚步。最终自主学习系统会被封装的越来越简单，大部分人不需要做细节方面的事，但需要了解”如果有这方面的数据可以学习的话，或许我们可以做到”。
- Waston本质上是IBM制造的开放域电脑问答（Q&A）系统， IBM介绍时说“Waston是一个集高级自然语言处理、信息检索、知识表示、自动推理、机器学习等开放式问答技术的应用”，并且“基于为假设认知和大规模的证据搜集、分析、评价而开发的DeepQA技术”。虽然采用了深度学习中一些技术如迁移学习 (Transfer Learning)来解决一些问题，但与AlphaGo不同，它并不是完全采用深度学习技术的人工智能。它的主体思路并非深度学习，而是更接近心智社会（Society of Mind）。
- Watson成功的基石是两个技术产物：搭建Q&A系统的可拓展软件架构DeepQA和基于多核算法技术的快速进步以及集成思想的AdaptWatson方法论。
    1. DeepQA：此架构以”处理流程”的形式定义了分析问题的各个步骤，像环形办公室走廊，每一间办公室都做着自己特殊的工作，允许多重实现来产生多个结果。每一间办公室的工作都当做大规模并行计算的一部分而单独进行。系统在基于对问题和类型的不同理解上对多个不同的资源进行搜索，返回多种候选答案，任何答案都不会立即被确定，因为随着时间推移系统会收集到越来越多的证据用来分析每一个答案和每一条不同的道路，之后系统用几百种不同的算法从不同的角度分析证据得出上百种特征值或得分，这代表着在某一特定维度上一些证据在多大程度上支持一个答案，每个答案的所有特征值或得分综合为一个得分表示该答案正确的概率。系统通过统计学机器学习方法对大量数据集进行学习来确定各个特征值的权重，最终确定答案的得分排名。

2. AdaptWatson：项目以最初的框架为基础，逐步改进核心算法，测量相应结果，然后再提出新的思想提高效率，不断重复这个过程来迭代。通过对多核心算法组件的研究、开发和升级使理解问题、搜寻答案、收集证据、给证据和答案评分等计算过程的速度变快，加快整个项目的迭代速度。

- 对一个句子进行语法解析是一个几乎完全解决的问题，Watson的问题理解模块只是在语法解析的基础上加入一个定型词（lexical answer type, LAT）的概念表示线索中能够推导出指示事物类型的单词或短语，用于帮助找出问题的指向。但即使语法分析十分精确也很难从结构化的数据库找到相同的表示，因为语言的表达具有多样性。所以Watson的答案发掘模块不假设自己能服从任何预设模型、文本类型集或数据库表来完全理解问题，而是首先生成一个广泛的备选答案集合，每个答案作为一个独立假设的初始点，产生一个独立推理流程尝试发现和评估所有有利的证据获得正确率。
- DeepQA技术现在还只是根据一个问题通过搜索和量化评估给出一个确定的答案，局限性非常大。未来我们希望DeepQA技术能够参与到与用户的互动中来在大量的非结构化内容支持下为用户提供决策，过去的决策工具需要根据目标领域人工编写形式逻辑模型，这种形式结构无法高效的跟上源知识的增长和领域的切换。我们需要系统在知识模板、搜索查询和基于深度学习的自动生成三者之间找的一个平衡点，在这个基础上还能考虑与用户互动的整个场景，低代价的消耗和分析自然语言资源获得最合适的答案。
- 微信在中国是智能对话最好的宿主平台，因为微信已经在中国普及，且由于app质量低劣和"墙"的存在难以获取高质量的信息，所以大部分人每天通过订阅号和服务号获取新闻和快捷服务，使用率极高。

C. 存留问题

- 未来对话系统在问题进入时会收集越来越多的用户信息比如情绪、手势、声纹，怎么平衡这些信息对生成答案时的影响力度？
- 增强对话系统的"人性"需要用户建模、记忆推理等一些新的模块加入，如何更好的完成这些功能？

D. 接下来调研方向

- 业界不同对话系统技术比较
- 多轮问答上下文引入
- 对话系统新的技术挑战

返回目录

2016-07-26 对话系统

A. 近期阅读文章

1. 巨头们都很重视的聊天机器人，你不进来看看吗？
2. 神经网络
3. Sequence to Sequence Learning with Neural Networks
4. [译] 理解 LSTM 网络
5. Attention based model 是什么，它解决了什么问题？ - 回答作者: Tao Lei

B. 个人理解

- 对话系统分为开放域(open-domain)和面向具体任务(task-oriented)两种。开放域对话系统要求AI具备各种知识，能完成多项任务，同时具有社会性（友好度、自觉性、幽默感等），现在的技术短期难以实现，市面上的这类系统还仅限娱乐。而任务导向的对话系统只需专注完成一项任务，实现相对简单，但是具体任务中数据规模小，难以通过data-driven的方式实现，需要人为整理知识库或对话模板，耗费大量人力，且难以向平行任务领域迁移。
- 四种对话生成的解决方案：
    1. 直接根据context生成对话，通过data-driven的方式预测生成文字，比较流行的框架是sequence to sequence + attention，RNN（Recurrent Neural Networks）加入了循环使得信息可以从当前步传递到下一步，所以是对于这类数据最自然的神经网络架构。
    2. 根据context从收集好的候选列表中选择最合适的回答，数据集准备麻烦，而且在实际应用中不好被借鉴。
    3. 基于规则或基于模板，从context中抽取关键value填充模板，适合解决具体的单一任务，向其他任务迁移时需要重新定制模板和规则。
    4. 基于问题，从知识库中找到与context最相近的问题，取出对应的答案，适合做不需要演进回答的对话系统。（刚开始学做微信公众号的时候自己用公众号实现了一个这样的，用lucene对知识库创建一个索引，找到匹配评分最高的问题的答案返回）

- 对话系统需要对用户建模，分析对话日志等隐式反馈信息，针对不同用户进行回答。
- 传统神经网络不具备记录上一层信息的功能，而RNN则加入了循环，使信息可以持久化。但RNN存在长期依赖（Long-Term Dependencies）问题，相关信息和当前预测位置之间的间隔变大时RNN会丧失学习到连接很远的信息的能力。
- LSTM（Long Short Term Memory）网络在RNN的基础上增加了细胞状态（Cell），在整个链上运行，只有一些少量的线性交互，信息在上面流传保持不变会很容易，避免了长期依赖问题。LTSM在链式形式的重复模块（block）中加入"门"结构来去除或者增加信息到细胞状态中，控制需要的信息。
- 加入注意力（Attention）会让LSTM效率更高。"Attention"的意思就是选择恰当的context而不是全部，并用它生成下一个状态。
- 实现一个任务导向的对话系统是更切实可行的，比如一个QA系统或判断句子正误的系统，在data-driven的解决方案中增加相关的knowledge sources、信息检索或是自动推理等一些模块组成一个固定框架，得到的效果会好很多。
- 对话系统难以评价，需要从用户体验中得到评价。

C. 存留问题

- 机器学习基本知识和思想不熟悉
- 神经网络是一个量化评估的工具？怎么应用与实现？
- 如何选择恰当的context当做"Attention"？强化学习（Reinforcement Learning）？

D. 接下来调研方向

- IBM Waston相关，IBM公司制作的电脑问答（Q&A）系统，在美国的电视问答节目Jeopardy!（危险之旅！）上击败了两名人类冠军选手，一战成名。
- Facebook Messenger Platform相关，FB公司有35亿活跃用户的社交数据，是最卖力宣传人工智能的公司。
- 机器学习基本知识和思想

返回目录