

SACC 第八届中国系统架构师大会
2016 SYSTEM ARCHITECT CONFERENCE CHINA 2016

架构 创新之路

深度学习在自然语言处理中的应用

- 李成华

目录

- 1 自然语言处理
- 2 深度学习
- 3 深度学习在自然语言处理中的应用



自然语言处理

Natural Language Processing

SACC 2016 第八届中国系统架构师大会

SequeMedia
盛拓传媒

IT168
中国IT门户网

ChinaUnix
中国Unix用户社区

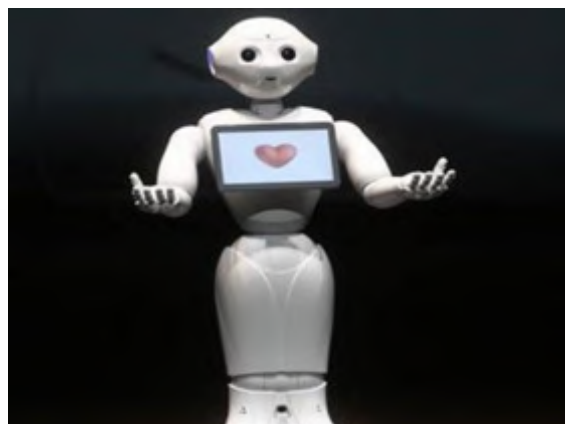
ITPUB
www.itpub.net

■ 自然语言的概念

● 什么是自然语言

- 语言是人类交际的工具，是人类思维的载体
- 人造语言：编程语言，包括C++，BASIC等
- 自然语言：
 - 形式：口语、书面语、手语
 - 语种：汉语、英语、日语、法语...

● 语言学是研究语言规律的科学



■ NLP 的应用

● NLP应用前景

- 据统计，日常工作中80%的信息来源于语言，处理文本的需求在不断增长
- 文本是人类知识最大的存储源，并且文本的数量在不停地增长
- 电子邮件、新闻、网页、科技论文、 用户抱怨信

● NLP典型应用

- 智能搜索引擎、自动问答、信息获取、语义网
- 语音识别，文字识别、输入法
- 机器翻译，自动文摘，跨语言检索
- 文本分类、文本聚类、文本分析（结构、内容、情感）、文本挖掘（主题跟踪：人物跟踪，企业跟踪）、文本过滤

■ 自然语言处理的层次

语音分析：

从语音流中区分出一个一个声音单元----音素

词法分析：

从句子中切分出单词、找出词汇的各个词素，确定单词的词性、词义等。

句法分析：

对句子和短语的结构进行分析，找出词、短语等的相互关系及在句子中的作用等。

语义分析：

识别一句话所表达的实际意义。

语用分析：

研究语言所在的外界环境对语言使用所产生的影响。

■ 语义与语用

- 同一词语在不同的“语境”中具有不同“语义”

例如：中国奥运史上十大女杰的精彩“转身”

- 病毒

计算机领域：计算机病毒

医学领域：生物学病毒

■ 内容层的信息处理

形态丰富的语言（inflecting language）：处理难

形态不丰富的语言（analytic language）：处理更难

中文	英文
学生们 都 来了	All the students are here.
李教授 都 来了	Even Prof. Li is here.
编辑 工作很难	Editing is very difficult.
怎样成为一个好的 编辑	How to be a good editor ?

■ 机器能够理解人的语言吗？

可以，但有难度！

什么是理解？

结构主义：机器的理解机制与人相同（白盒）

问题：人类语言理解机理尚未清楚

功能主义：机器的表现与人相同即可（黑盒）

图灵测试

如果通过自然语言的问答，一个人无法识别和他对话的是人还是机器，那么就应该承认机器具有智能

■ 理解自然语言的准则

给计算机输入一段自然语言文本，计算机能

问答(question-answering)

机器能正确地回答输入文本中的有关问题；

文摘生成(summarizing)

机器有能力产生输入文本的摘要；

复述(paraphrase)

机器用不同的词语和语句复述输入文本；

翻译(translation)

机器把一种语言(源语言)翻译为另一种语言(目标语言)

■ 自然语言处理的研究目标

弱人工智能目标：

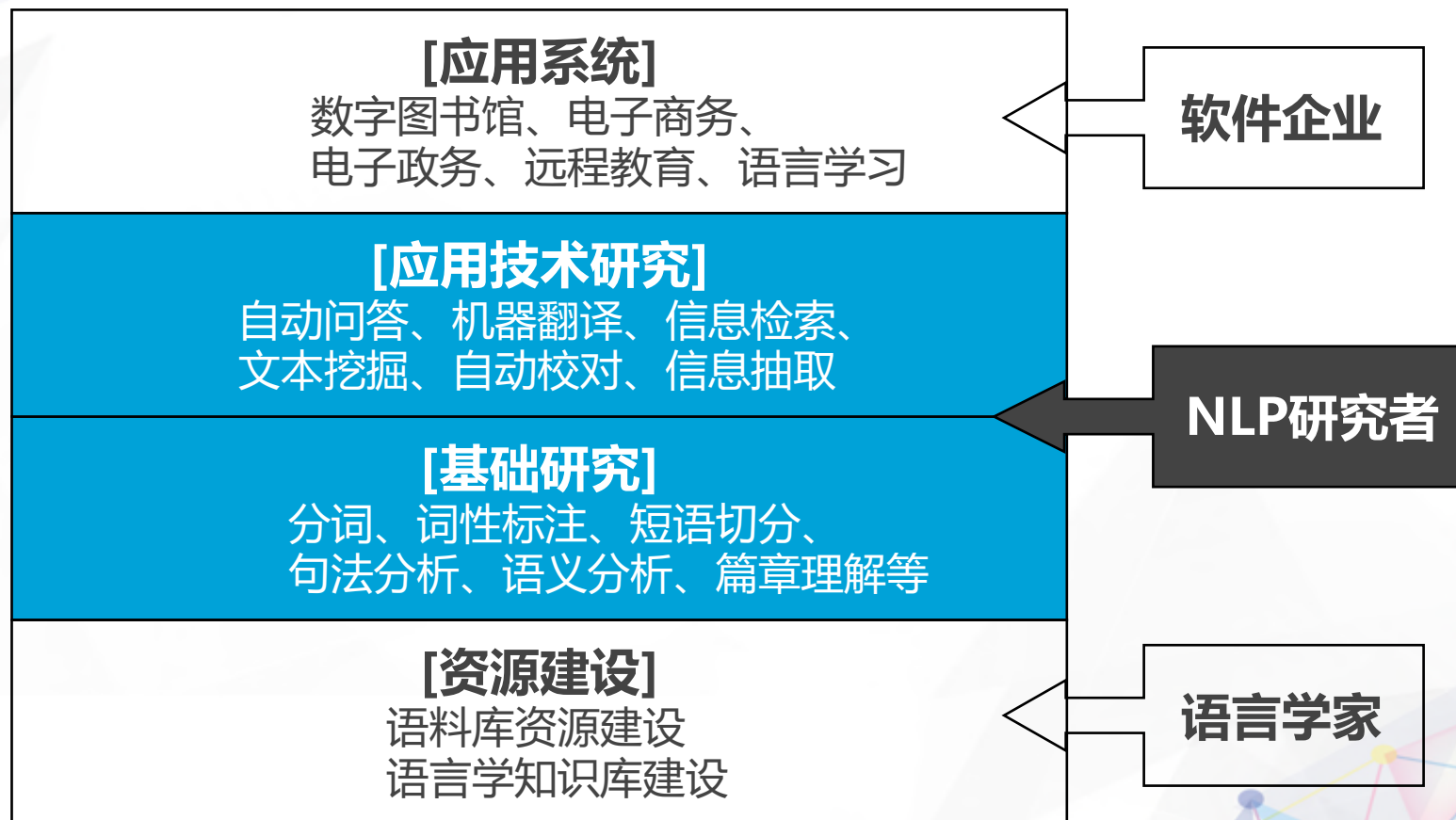
建立一个足够精确的语言数学模型使计算机通过编程来完成自然语言的相关任务。如：听、读、写、说，释义，翻译，回答问题等；

强人工智能目标：

让用户能通过自然语言与计算机自由对话；



■ NLP研究内容



■ 自动分词

主要分词算法:

1. 基于词表的分词-最大匹配
即“长词优先”原则，来进行分词
2. 基于统计的分词

分词的难点

歧义字段处理

未登入词处理: 人名识别, 地名识别, 译名识别, 新词识别

■ 分词歧义

例：南京市长江大桥
南京|市长|江大桥
南京市|长江|大桥

例：我们研究所有东西

- 我们--研究所--有--东西（交叉歧义）
- 我们--研究--所有--东西
把手放在桌上
- 把--手--放在--桌上（组合歧义）
- 把手--放在--桌上

■ 分词歧义

新词

名词：斑竹、大虾、面瓜、菜鸟、美眉、陶吧、911、白骨精
动词：打的、埋单、买单、给力
形容词：酷、小资、爽

命名实体

时间、地名、人名、组织机构名
商标、公司名、电话号码、电子邮件地址等等

合成词，简称，派生词

三个代表、十六大
牛市、熊市
微博

■ 词义消歧

机器翻译中最难的问题之一是词义的二义性（歧义性）问题。

- 比如 Bush 一词可以是美国总统的名字，也可以是灌木丛。

一个实用的方法是使用互信息。

- 互信息(Mutual Information)是信息论里一种有用的信息度量，它是指两个事件集合之间的相关性
- 首先从大量文本中找出和总统布什一起出现的互信息最大的一些词，比如总统、美国、国会、华盛顿等等，当然，再用同样的方法找出和灌木丛一起出现的互信息最大的词，比如土壤、植物、野生等等。有了这两组词，在翻译 Bush 时，看看上下文中哪类相关的词多就可以了。



深度学习

Deep Learning



■ 什么是深度学习?

深度学习：

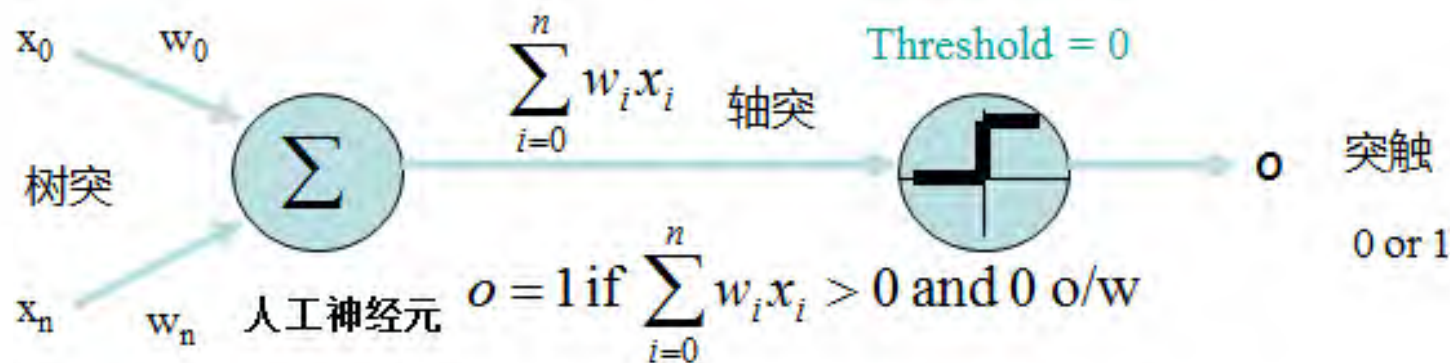
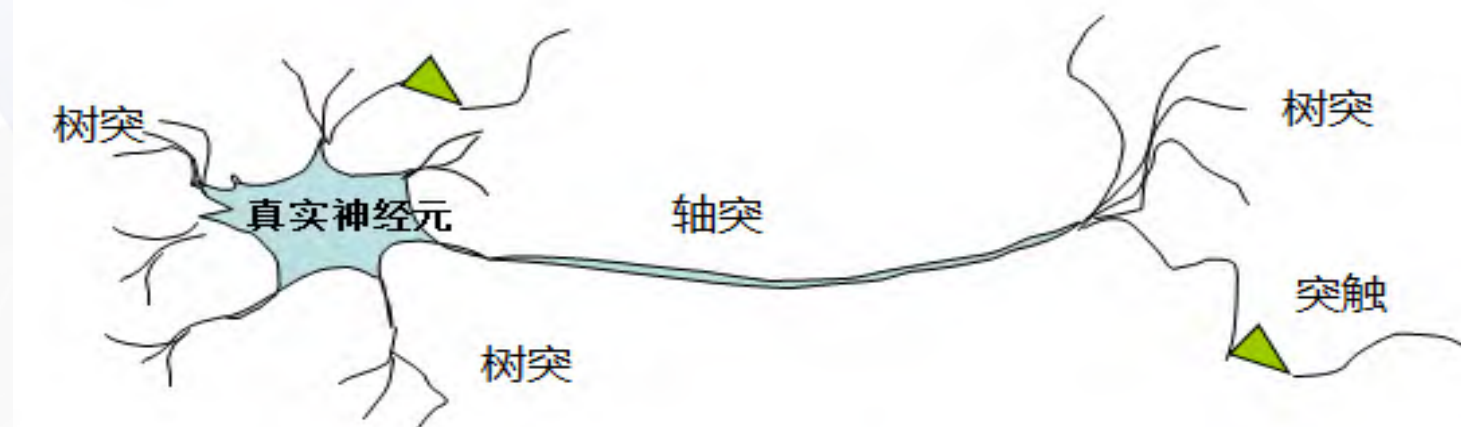
一种基于无监督特征学习和特征层次结构的学习方法。

本质：

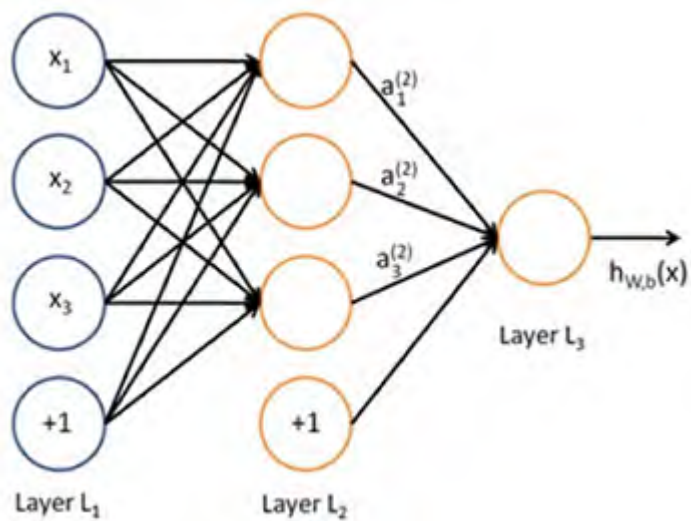
通过构建多隐层的模型和海量训练数据，来学习更有用的特征，从而最终提升分类或预测的准确性。

含多隐层的多层感知器就是一种深度学习结构。

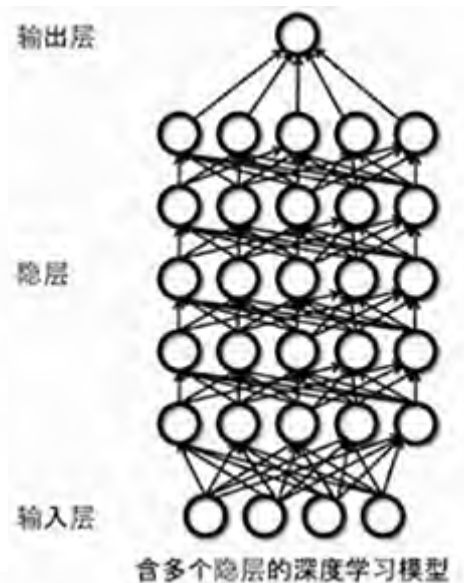
■ 真实神经元 VS 人工神经元



■ 人工神经网络 VS 深度学习



传统神经网络



深度神经网络

深度学习是模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，处理数据。

深度学习发展历程

1986, BP ,
Nature
Rumelhart,
Hinton,
Williams



2011
PhD
candidate
of Hinton,
MSR



2015
Deeper Network
ImageNet
Classification
Speech



2006
DBN
Science
G.E.
Hinton

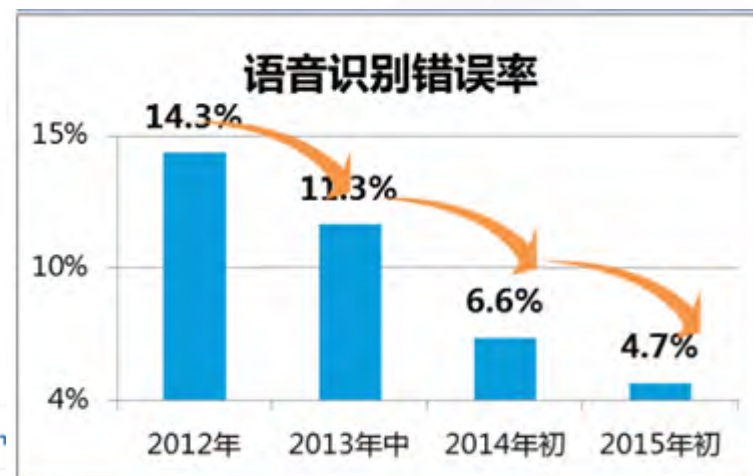
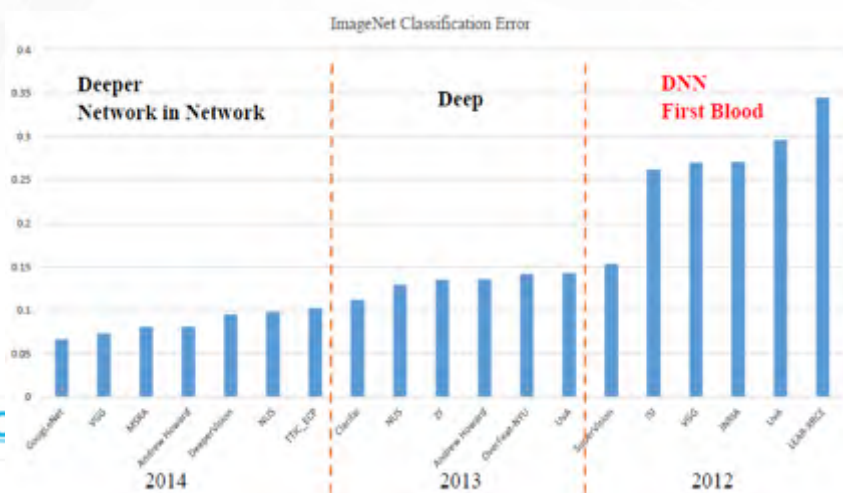
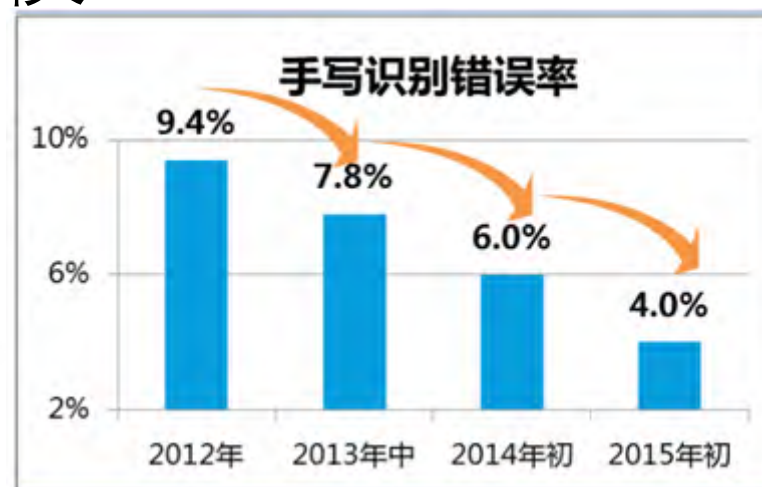
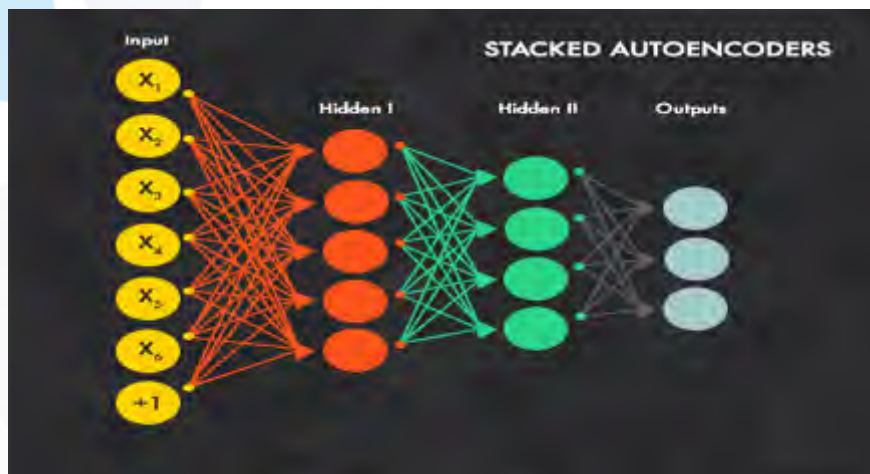


2012 DNN
ImageNet
Classification



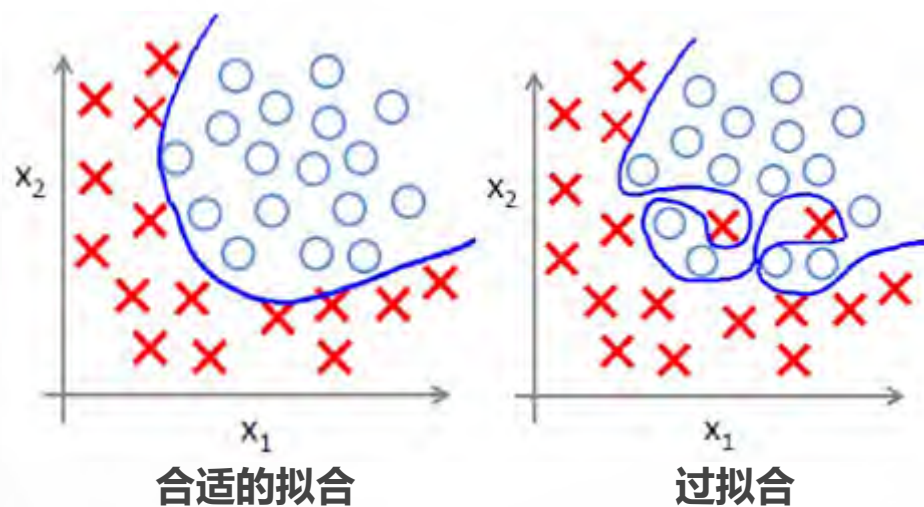
深度学习-2013年十大突破性技术之首

500亿美元的市场规模



■ 传统学习方法的缺点

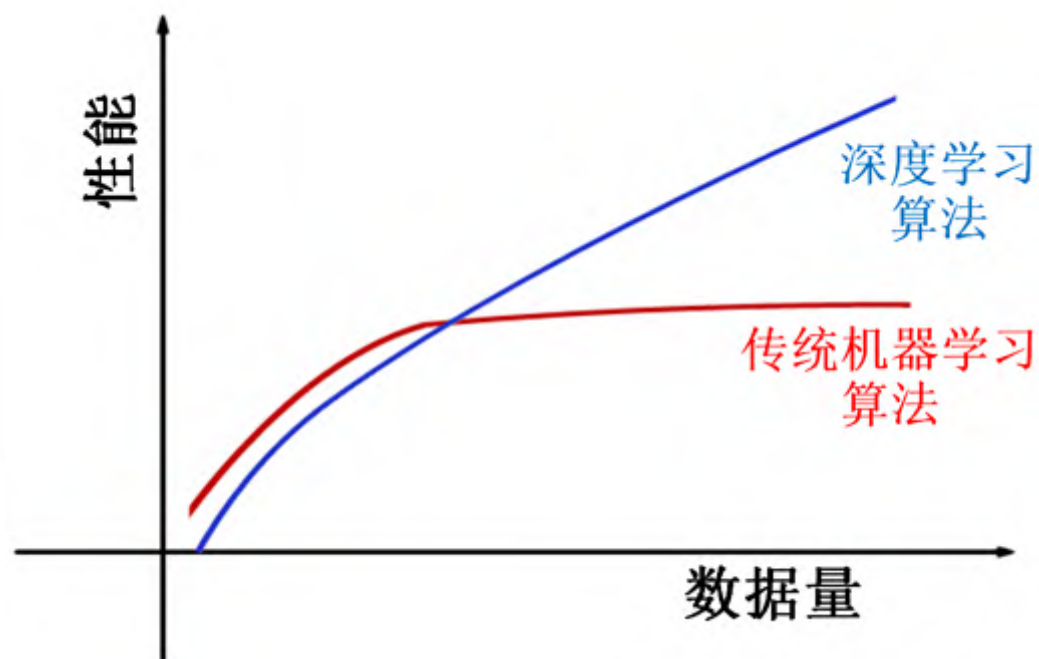
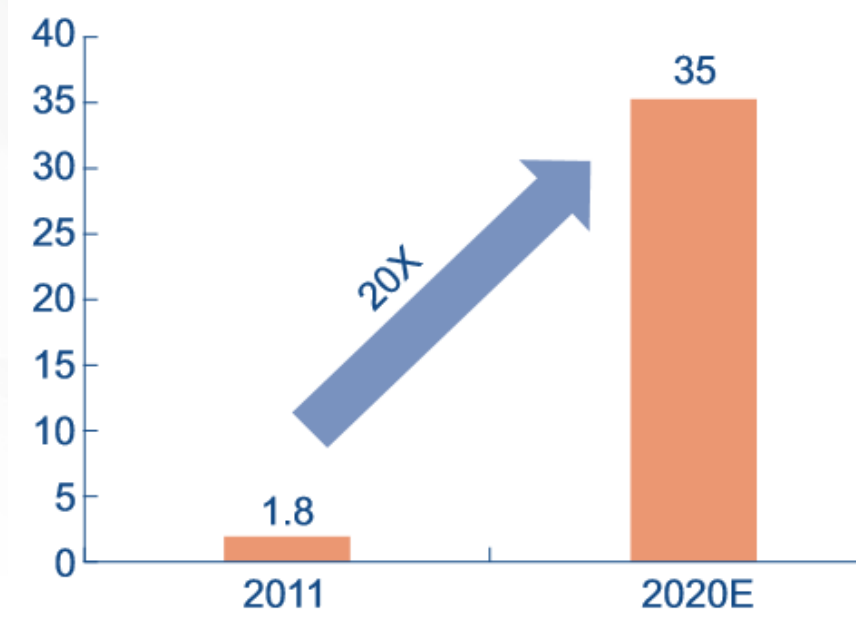
过拟合 (Overfitting)



过拟合往往源自特征空间分布过于复杂，使得模型的复杂度提高，泛化能力较差。

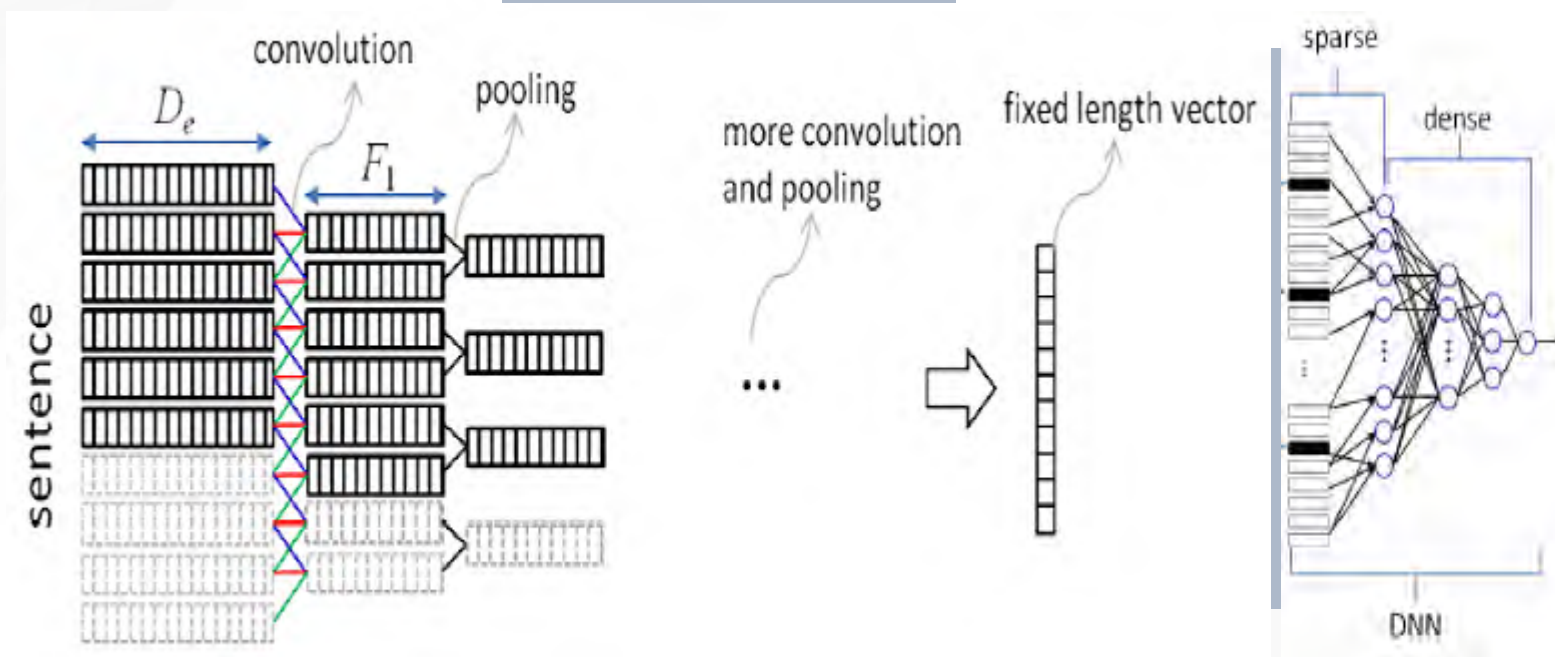
深度学习优点

全球数据总量 (ZB)



深度学习在自然语言处理

DCNN (Deep Convolutional Neural Networks) 深度卷积神经网络



■ 向量空间模型 Vector space model

1

- 维度高
- Time consuming

2

- 一词多意和多词一意
- Mis-classification

3

- 缺乏词和词之间的相关性
- Poor performance



“Tank”

“Car” ?



?

“Automobile”



■ One Hot

One Hot 表示在传统 NLP 中很常用

"dog" = $[1, 0, 0, \dots, 0]$

"cat" = $[0, 1, 0, \dots, 0]$

"the" = $[0, 0, 0, \dots, 1]$

Similarity(dog, cat) = 0

■ Word Embedding

词向量：单词的分布向量表示 (Distributional Representation)

"dog" = [1, 0, 0.9, 0.0]

"cat" = [1, 0, 0.5, 0.2]

"the" = [0, 1, 0.0, 0.0]

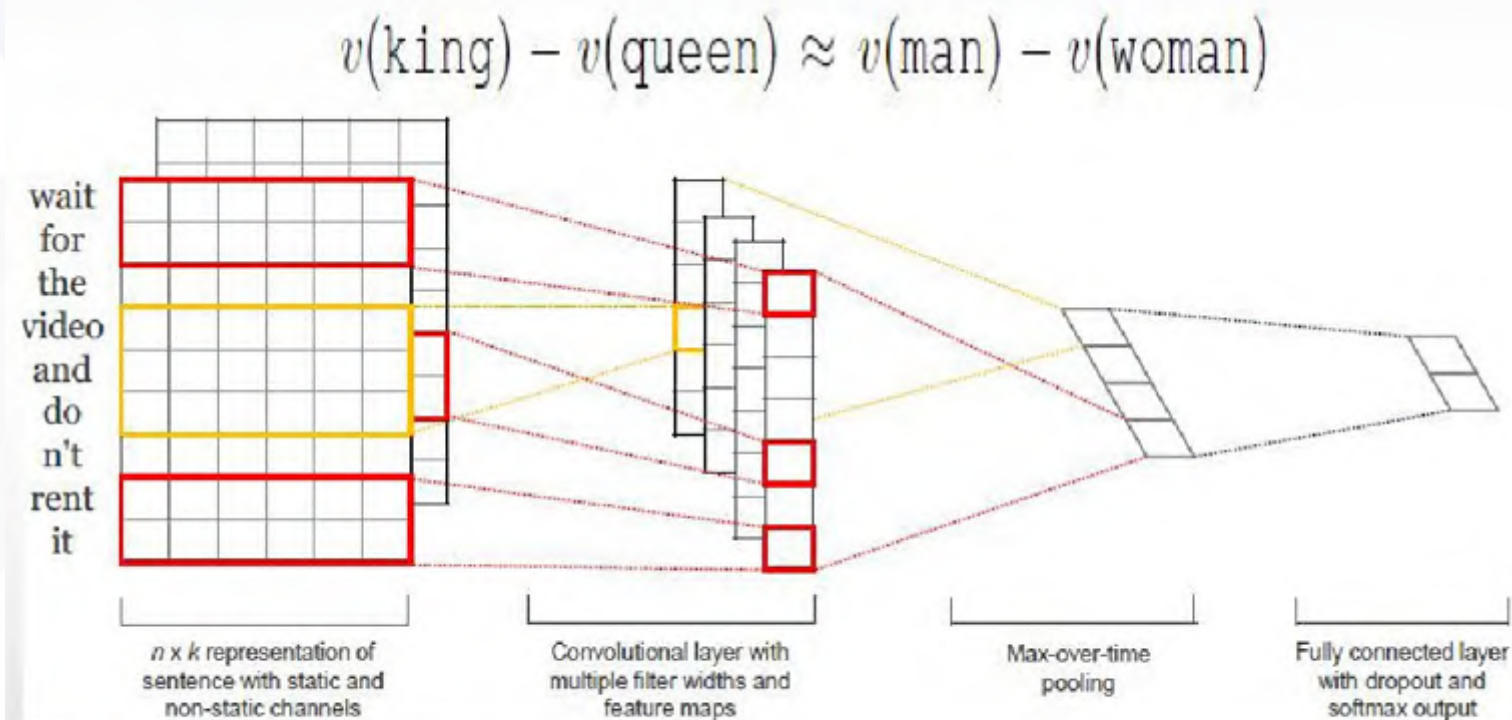
Similarity(dog,cat) > Similarity(dog,the)

Similarity("the dog smiles.", "one cat cries.")

词向量表征了单词使用上下文中的句法语义特征

One-Hot的字面匹配到DR的语义匹配

word2vec原理



word2vec实战

训练数据集：经过分词后的新闻数据，大小184MB

查看“中国”，“钓鱼岛”，“旅游”，“苹果”几个词语的相似词语如下所示

请输入词语<exit退出>:中国

大陆	0.66763467
中共	0.57856727
共产党	0.56305367
解放军	0.55761635
台湾	0.5368497
反攻	0.5271177
日本	0.5103535
王文莹	0.49295437
内地	0.48557448
对岸	0.48428434

请输入词语<exit退出>:钓鱼岛

钓鱼台	0.6219264
钓岛	0.6123347
南海	0.6018163
领土	0.51753837
领海	0.4928774
岛屿	0.4853142
舰队	0.47854927
渔权	0.47229362
主权	0.46729872
东海	0.4613399

请输入词语<exit退出>:旅游

观光	0.65619475
景点	0.60212
陆客	0.59477097
旅行	0.5677106
游憩	0.557839
赏樱	0.5571045
游玩	0.52199984
观光客	0.51974636
行程	0.51943743
参观	0.5077874

请输入词语<exit退出>:苹果

三星	0.7224437
微软	0.7101249
Apple	0.66682446
iPhone5	0.62071097
Google	0.597368
iPadmini	0.5609188
新机	0.559093
库克	0.5589176
宏达电	0.555409
产品	0.55437565

word2vec原理

向量加减法

"中国+北京-日本", "中国+北京-法国", "家庭+孩子-学校"

请输入词语<exit退出>:中国,北京,日本

中国 + 北京 - 日本 =

东京 0.6188478

外海 0.4843038

广东 0.45942163

安倍 0.44709134

参拜 0.44158116

派出 0.43384194

釜山 0.4310615

海域 0.42759195

湖北 0.4272585

官邸 0.42487407

请输入词语<exit退出>:中国,北京,法国

中国 + 北京 - 法国 =

巴黎 0.7990697

伦敦 0.7614885

义大利 0.7026581

纽约 0.6534368

德国 0.6426668

首都 0.6354907

英国 0.6339937

西班牙 0.6326488

柏林 0.628184

加拿大 0.6255576

请输入词语<exit退出>:家庭,孩子,学校

家庭 + 孩子 - 学校 =

老师 0.948024

国中 0.76681674

学生 0.73207045

高中 0.7140167

小孩 0.70768154

小朋友 0.692751

上课 0.6889297

校方 0.687891

家长 0.66356415

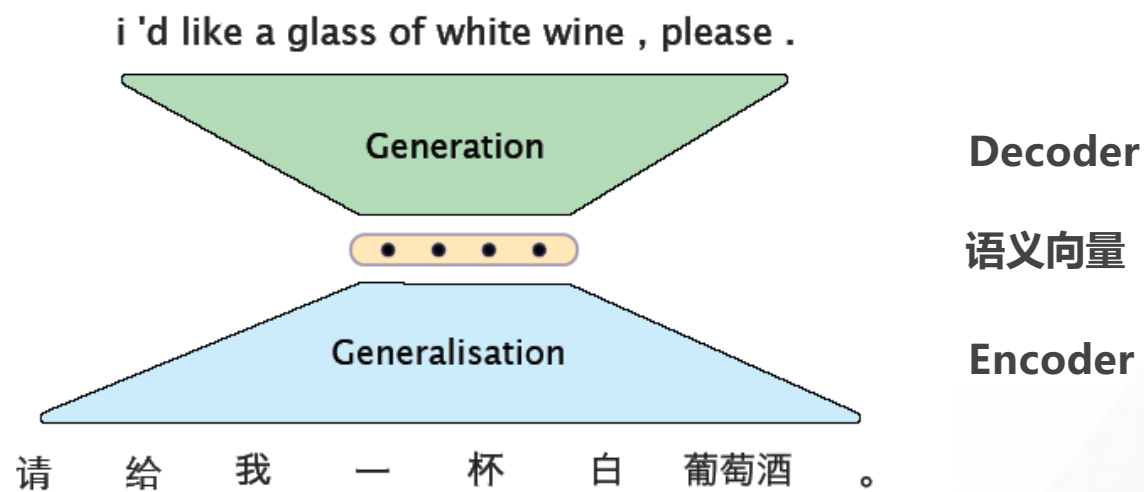
高职 0.65407264

■ 深度学习的局限

1. **适用于大数据**，由于网络结构复杂，小数据集往往会造成训练结果过拟合。
2. **可解释性不强**，即便是一流的学者也很难对效果超群的深度学习算法在具体问题上给出具体解释。这有可能成为产品迭代过程中的阻碍。
3. **理论尚不完善**，大多数研究仍集中在调整参数和结构。
4. **计算量大**，算法调整费劲。

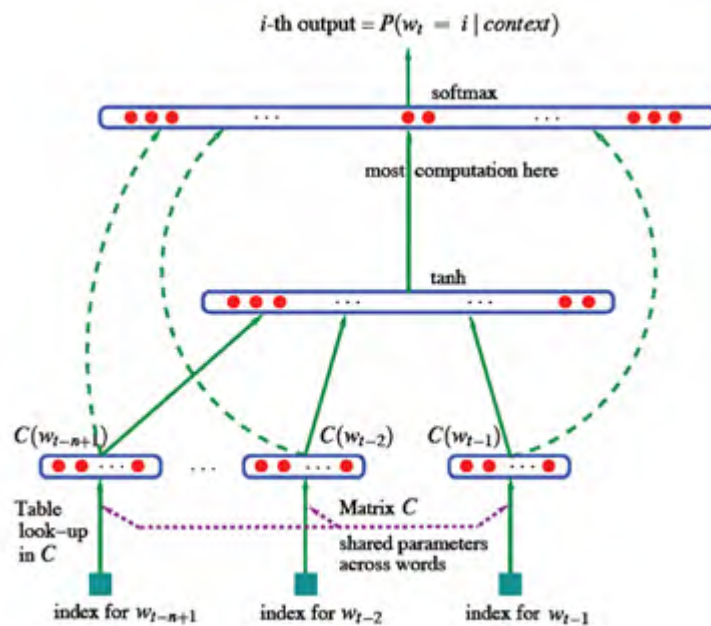
深度学习在自然语言处理中的应用

■ 机器翻译（通用模型）



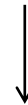
最常见的通用模型：Encoder-Decoder Model

■ 机器翻译 (Encoder-Decoder具体例子)



网络结构

语言模型



$$P(T|S) \approx \prod_{i=1}^{|T|} P(t_i | t_{i-1}, \dots, t_{i-n+1}, s_i)$$

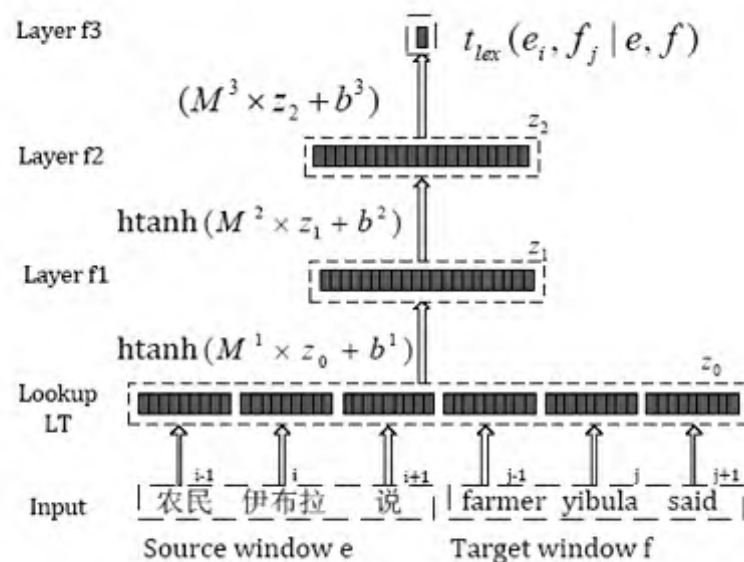
翻译模型



■ 机器翻译-很多地方可以运用到深度学习

单词对齐
短语对齐
短语重排序
语言模型
翻译模型
联合模型
翻译结果重排序

.....



单词对齐

人工客服or智能客服？

☹️ 人工客服

无法数据分析

无法控制时长

高成本

高峰期排队

响应慢

易出错

知识培训频繁

不能全时段覆盖

低效率

回复口径不一

人员流失率大



😊 智能客服

顶级语义分析

7*24小时工作

高智商

1↔100用户

高情商

极速响应

大数据分析

提升客服价值

多平台支持

智能知识库管理

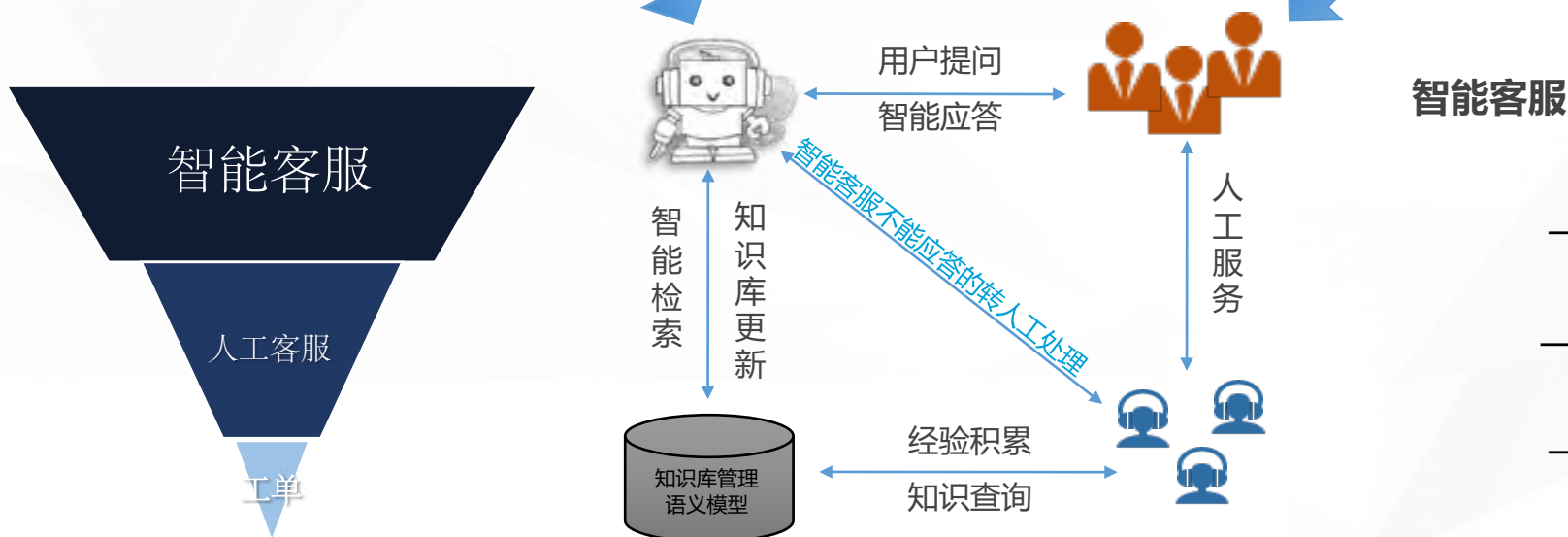


人工客服&智能客服

传统客服业务流程

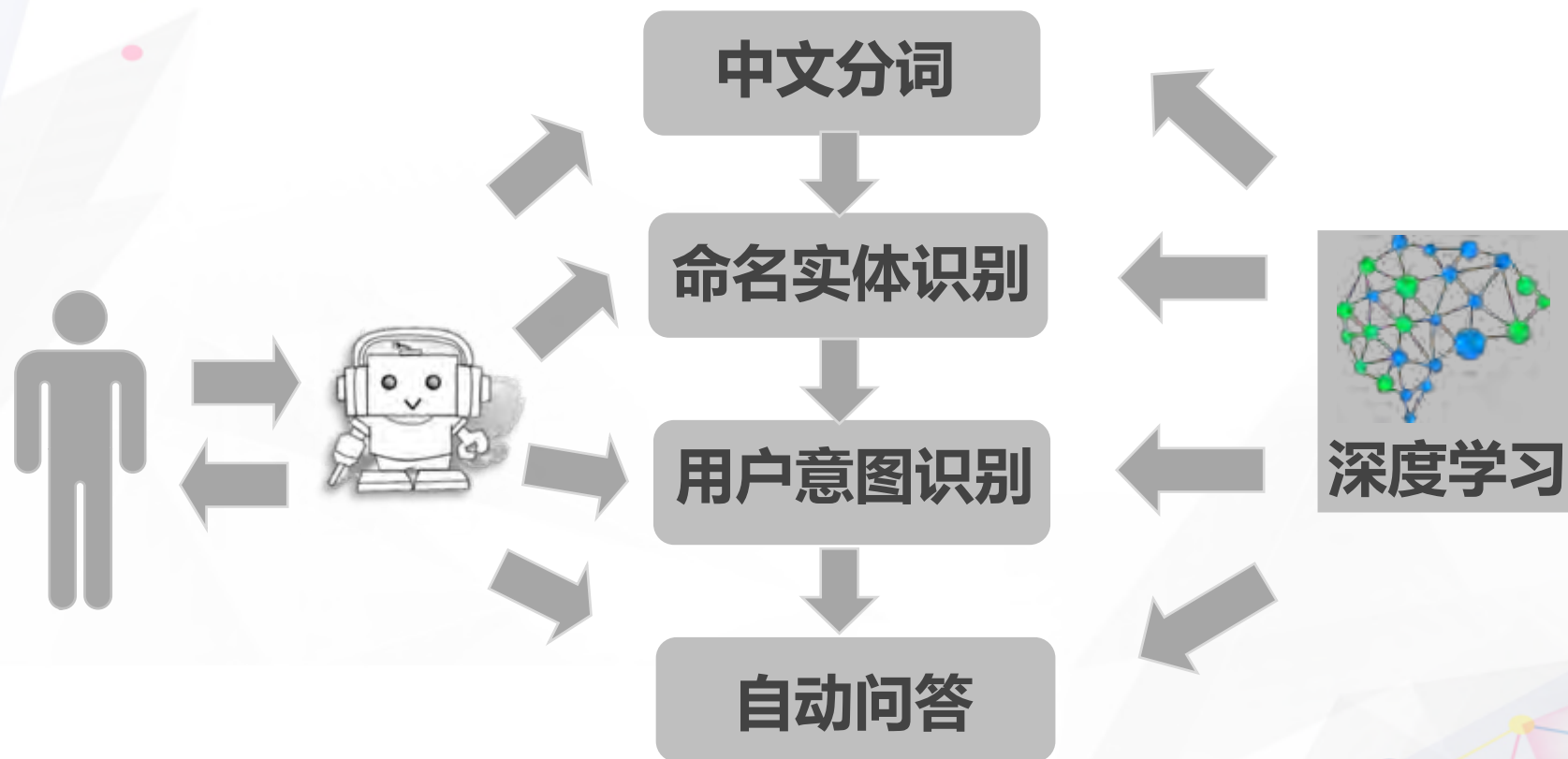


智能客服业务流程

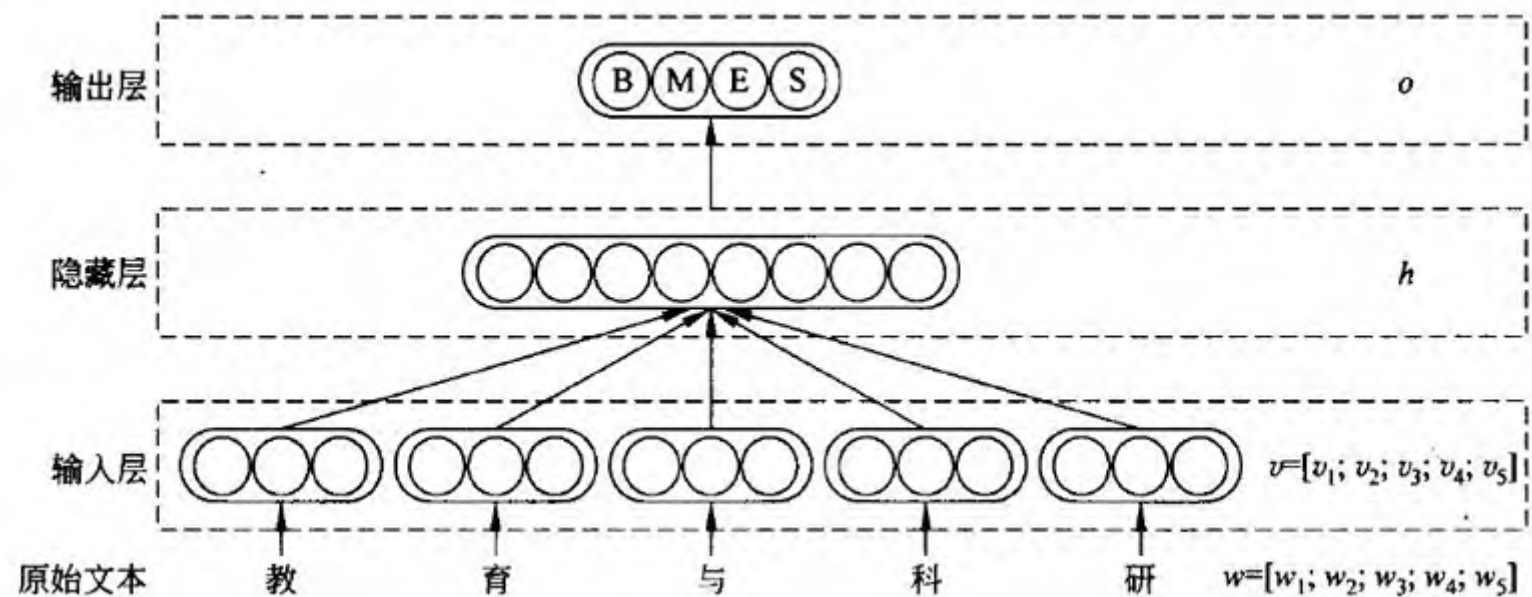


- 秒级响应
- 处理85~90%的问题
- 准确率超过85%
- 成本80%↓

■ 深度学习在客服机器人上的应用流程



深度学习用于中文分词



■ 深度学习用于中文分词

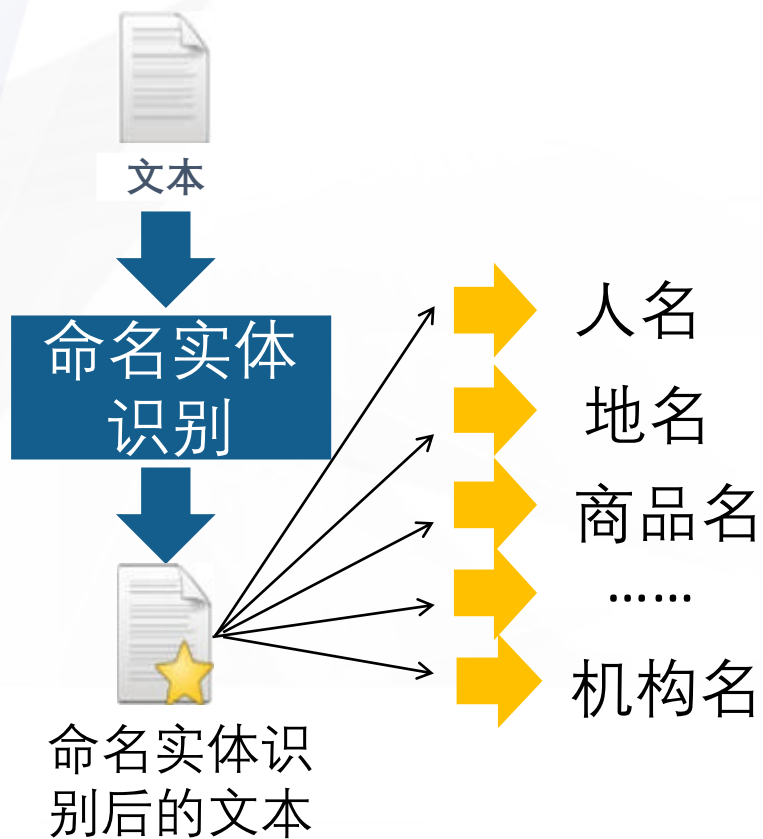
中文分词思路：

基于字的Word Embedding + 三层神经网络 + BMES标记序列分类

类似思路同样可以套用到POS/NER/Parser等场景

这是利用Word Embedding解决NLP问题最直观的NLP应用思路

命名实体识别



- ✓ 对用户的输入进行识别
- ✓ 对识别后的命名实体进行抽取
- ✓ 更好地理解用户语言和意图

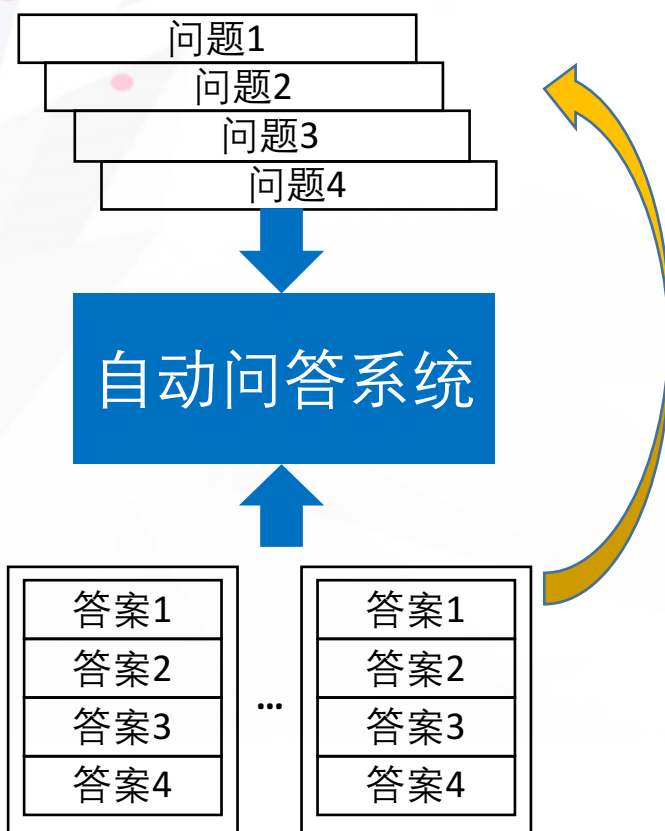
用户意图识别



意图识别是智能客服系统里面非常重要的一部分,它首先要通过用户的问话,知道用户想要做什么,只有意图识别对了,才能在相应的类别里面反馈用户的答案



自动问答



- ✓ 对用户的问题进行意图识别
- ✓ 确定问题的分类
- ✓ 答案匹配，抽取和排序
- ✓ 给用户反馈最佳答案和建议



答案匹配，候选
答案抽取和排序

智能聊天机器人



IBM Watson



Apple SIRI

Microsoft

微软二代小冰

“在亿万人之中，我只属于你。”

微软小冰



小 i 机器人



泰迪 机器人

如何选择
智能客服



- ❖ 特定领域下问题匹配的准确度
- ❖ 机器人转人工的策略，及灵活性、可扩展性
- ❖ 上下文理解能力（场景）
- ❖ 知识的积累和扩展性（智能管理）
- ❖ 用户意图识别的准确性
- ❖ 和其他业务系统可集成性
- ❖ 深度定制开发



2016 第八届中国系统架构师大会

SequeMedia
盛拓传媒

IT168
中国IT网

ChinaUnix
中国Unix网

ITPUB
www.itpub.net

STOR/AOR 融拓

大数据人工智能服务提供商

■ 问答实例



liuyong 5:12 PM

@taidibot 解释一下一石二鸟的意思



taidibot BOT 5:12 PM



liuyong 5:17 PM

@taidibot 李姓的起源



taidibot BOT 5:17 PM



liuyong 5:29 PM

@taidibot 红烧肉的做法



taidibot BOT 5:29 PM

@liuyong 1.五花肉切块2.五花肉凉水下过焯至变色备用,笋切大块。3.锅内少放油,放适量白砂糖,我崇尚少油、少糖,您可以按照自己口味适量添加。待糖冒泡,即刻下入焯好的五花肉。4.五花肉尽快翻炒,炒出糖色。5.逐渐上色后,下入姜片、小辣椒、花椒、八角、桂皮、香叶、丁香、草果一起翻炒至闻到作料的香气。6.加入热水(肉容易软烂,如果加凉水,肉不容易烂,而且肉里的油不容易煮出来。),加入生抽、老抽、腐乳汁、盐、鸡精、料酒,大火烧开,改小火炖煮。7.半小时后,筷子能够插进肉里时,下入春笋,继续中小火炖煮。8.经过1小时的炖煮,肉已经彻底软烂了,开大火把汤收至粘稠。9.好啦,肥而不腻,带着笋香的红烧肉闪亮出锅啦10.搭配个小青菜,美美的一顿饭11.肉烂笋香,保证你能吃下两大碗米饭。



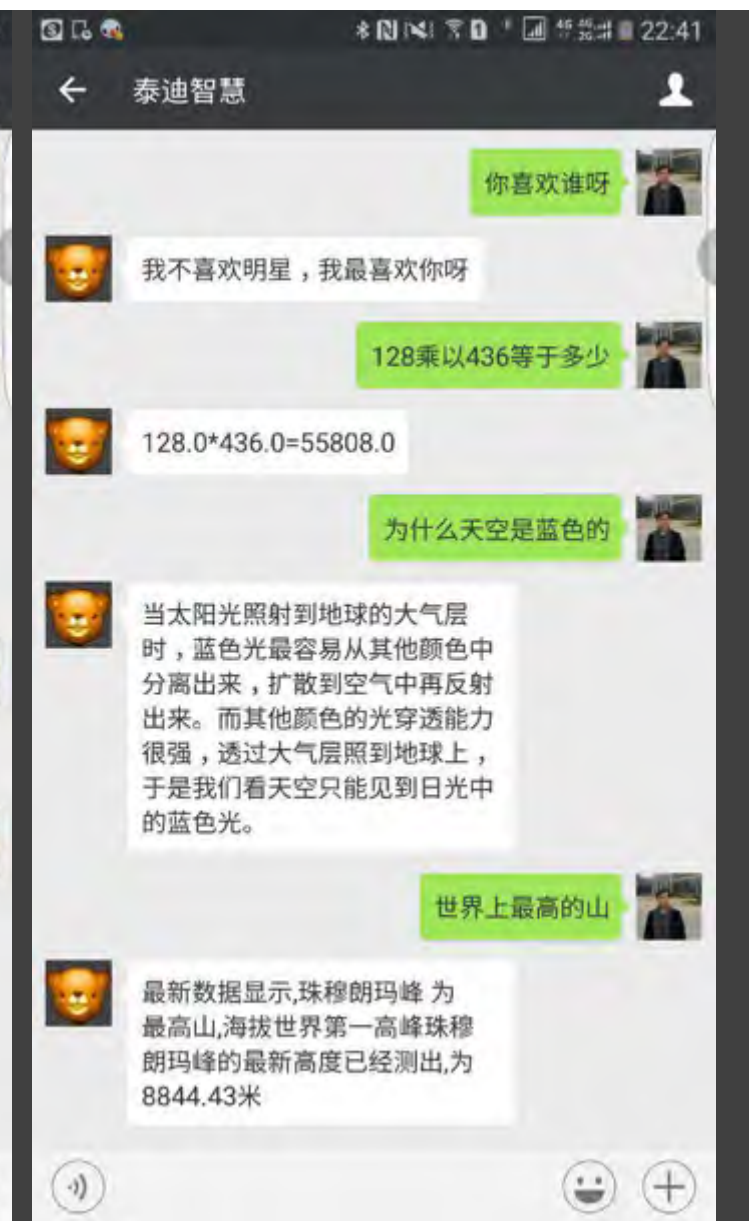
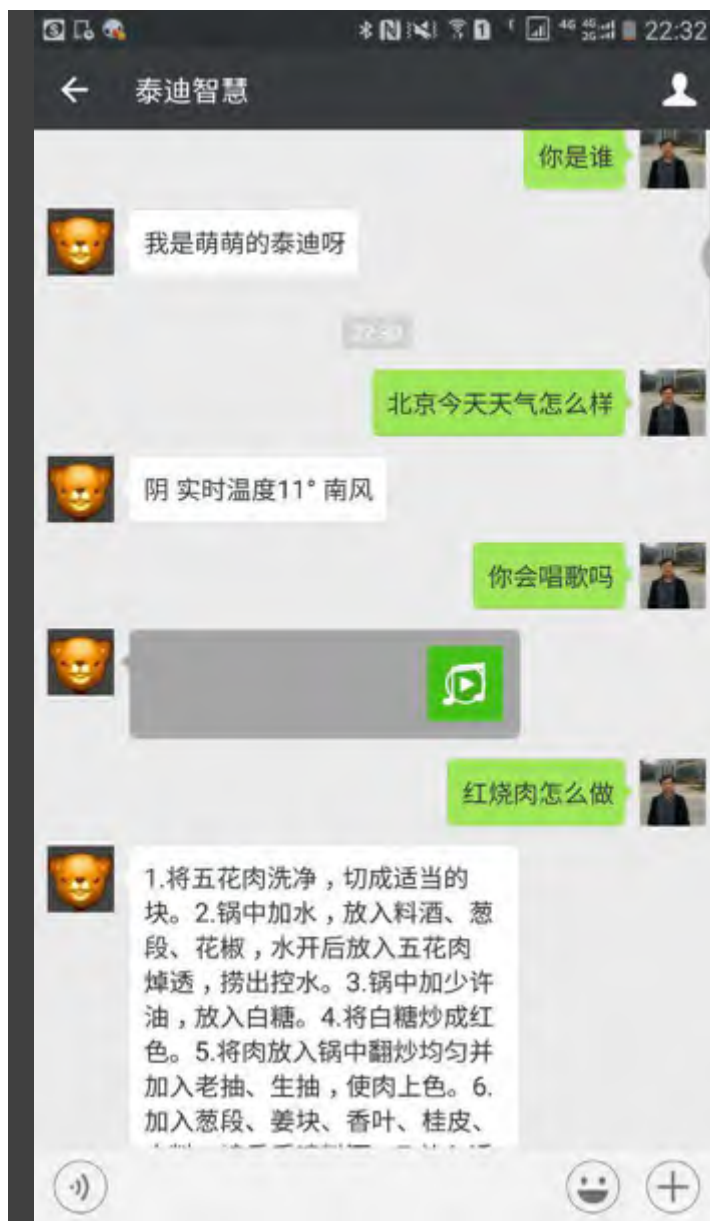
liuyong 5:29 PM

@taidibot 我想知道炸春卷的制作方法



taidibot BOT 5:29 PM

@liuyong 1.春卷皮买回来备用。2.猪肉洗净切丝,鸡蛋打散,葱洗净切碎备用。3.绿豆芽洗净备用。4.热油锅,先把鸡蛋炒好,用铲子压碎,盛出备用。5.热油锅,先把鸡蛋炒好,用铲子压碎,盛出备用。6.把肉丝放进去煸炒,炒至变色,加料酒和生抽炒一会。7.然后把绿豆芽放进去炒,炒至绿豆芽变软,熄火。8.熄火后,加入葱叶,盐,五香粉,香油。9.拌均匀,春卷的馅就做好了。10.取一个春卷皮,在春卷皮中间的三分之一放下拌好的馅。11.从少的部分折起来。12.然后在卷一折。13.把二头对折起来。14.然后继续卷,卷好为些。15.全部卷好。16.锅入适量的油,稍微多一点油,七八成热,下包好的春卷下去卷,先放有口的一面下去炸。17.炸至二面金黄即可。





SACC 2016 第八届中国系统架构师大会

SequeMedia
盛拓传媒

IT168
中国IT门户网

ChinaUnix

IT-PUB
www.it-ebook.com

李成华

lichenghua@stormorai.com

THANKS

SequeMedia
盛拓传媒

IT168
中国IT网

ChinaUnix

ITPUB
www.itpub.net