CCAI 2017
中国人工智能大会

深度学习与自然语言处理

中国科学技术大学

陈恩红 教授

深度学习应用于自然语言处理

- 随着深度学习在语音，图像上取得突破，大家开始将注意力转移到**自然语言处理**领域。

词汇蕴含识别

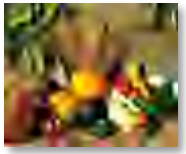
- **词汇蕴含识别**是**文本蕴含识别**的重要组成部分
  - 前提句：小明被一只狗咬了
  - 假设句：小明被一只动物攻击了
  - 如果我们可以知道：“狗”蕴涵“动物” “咬”蕴涵“攻击”

- 词汇蕴含识别的**应用**：单选题自动作答

You should take umbrella when you go out because it’s __ outside

A: rainy B: sunny C : cloudy D : overcast

- 词汇蕴含具有很大挑战



  - **词义**多样性：果实
  - **蕴含关系**多样性：因果关系、上下位关系
  - 组合方式多样性

我们提出了基于**深度神经网络**的词汇蕴含识别方法

试题难度预测

- 通过**题目文本信息**与**历史答题记录**，预测试题在新的考试中的难度。

英语阅读**文档**、**问题**及**选项**

从**历史**考试记录获取**试题难度**

Table 1: A toy example of test logs.

TestId	ExamineeId	QuestionId	Score
T_1	U_1	Q_1	1
T_1	U_1	Q_2	1
T_1	U_2	Q_1	0
T_1	U_2	Q_2	1
T_2	U_4	Q_3	1
T_2	U_5	Q_3	1
T_2	U_6	Q_3	0
...	...	...	...

试题难度预测

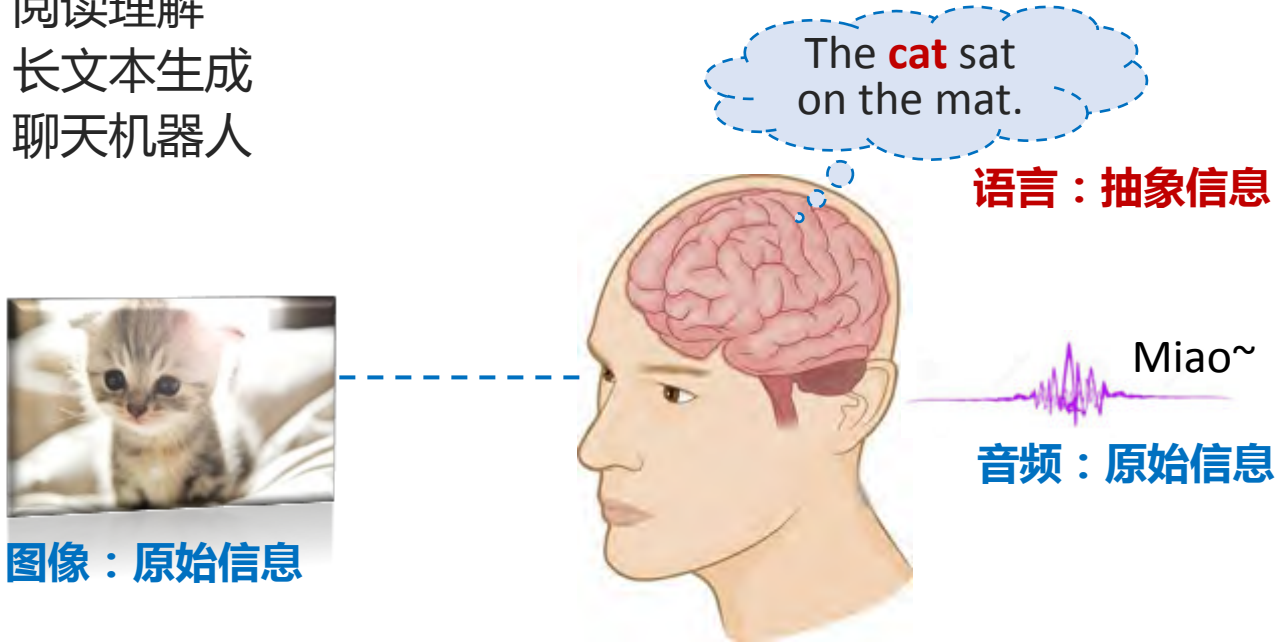
Table 2: Examples of question instances combined with test logs and question materials.

Difficulty (P)	QuestionId (Q)	TestId (T)	Text Materials					
			Document (TD)	Question (TQ)	Options (TO)			
0.4276	Q_1	T_1	Larry was on...	In what way...	His daughter had...	He had become...	His father...	His daughter...
0.4827	Q_2	T_1	Larry was on...	Why did Larry...	To protect himself...	To dive into...	To admire the...	To take photo...
0.5494	Q_3	T_1	Larry was on...	What can be...	Larry had some...	Larry liked the...	Divers had to...	Ten-year-old...
?	Q_4	T_2	Are you...	Why do people...	They eat too...	They sleep too...	Their body...	The weather...

我们提出了基于**深度卷积神经网络**的试题难度预测方法

深度学习在NLP领域应用尚未成熟

- 深度学习技术在自然语言处理领域已经**初步**取得了一些成果
- 然而相比其在图像、语音方面的应用已经趋于成熟，深度学习技术在NLP领域**尚处于探索阶段**
  - 阅读理解
  - 长文本生成
  - 聊天机器人

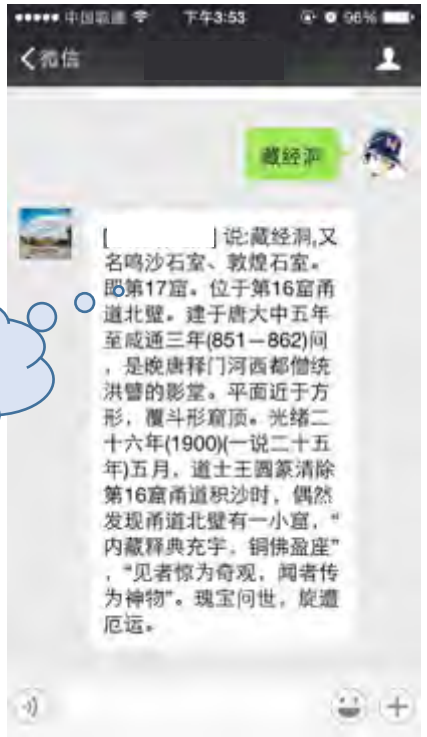


建模抽象信息

- 建模抽象信息需要的不仅需要**数据规模**庞大，还需要数据**多元化**
  - 大量多元化数据**难以获取**
  - 多元化数据**难以融合**
  - **领域知识**错综复杂

多元化情境
影响语义

问答系统需要
融入领域知识



知之为知之，不知为不知，**是知也**。

知道就是知道，不知道就是不知道，**这就是知道**。

数据量影响模
型拟合效果

规则到深度学习的过渡

- 针对数据量不足、类型不够丰富的挑战
  - 基于规则构建**原型系统**，从而**积累数据**
  - 随着数据量的提升与数据类型的丰富，**逐渐引入**深度学习技术

深度学习不能完全替代规则方法

- 即便拥有大量的多元化数据，我们也不能单独依靠深度学习处理一些较为复杂的问题
  - 数学试题回答

> 问题：甲、乙两车同时从A、B两站相对开出，第一次相遇离A站有90千米，然后各自按原速继续行驶，分别到达对方出发站后立即沿原路返回。第二次相遇时离A站的距离占AB两站全长的65%。求AB两站的距离。

  - 我们可以首先**通过深度学习的方法选取正确的规则**方法（确定题目类型，数学模型），然后运用相应的规则形式化的表示题目，进而求解

深度学习技术尚不完美

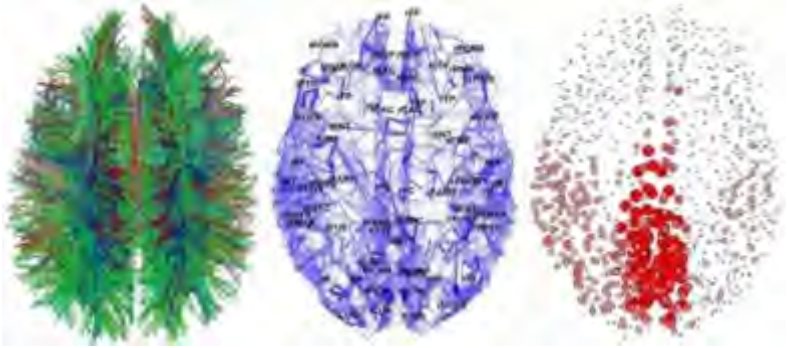
- 比如，神经网络会得出十分诡异的结果
  - 不正确,但也不是人类能理解的那种错误

Google 的神经网络画的诡异图画

神经网络和人脑并不相同

- 大脑中的一个单个神经元就是一个极其复杂的机器，即使在今天，我们也还不能理解它。而神经网络中的一个「神经元」只是一个极其简单的数学函数，它只能获取生物神经元复杂性中极小的一部分。所以，如果要说神经网络模拟大脑，那也只在不够精确的启发水平上是对的，但**事实上人工神经网络和生物大脑没什么相似之处**。——吴恩达

来自Google 、 Facebook 、纽约大学和蒙特利尔大学的研究者开发了一个神经网络系统,该系统认为**左图**是一只狗，而**右图**（仅在左图的基础上**略微改变了像素**）是一只鸵鸟。



最先进的大脑成像技术生成的有趣大脑视图

思考与展望

- 深度学习在NLP领域，如何能体现“深度”的思想？
  - 当前应用还说不上很Deep：以浅层的CNN和RNN为主。

- 深度学习在NLP领域，如何与其他技术结合？
  - 深度学习+强化学习：例如将商品购买情况作为reward，来引导基于深度学习的客服对话系统。
  - 深度学习+迁移学习：例如用迁移学习的方法将通用领域的情感分类模型迁移到微博情感分类。
  - 深度学习+多任务学习：例如将中译英，中译日等机器翻译任务进行多任务学习训练。

思考与展望

- 如何减少对标注数据的依赖，更好利用无监督数据？
  - 使用无监督语料上训练得到的词向量。
  - 对偶学习：机器翻译任务中，英译中和中译英互为对偶任务，可以利用中文和英文的单语语料，结合少量中英平行语料，利用对偶学习进行训练。

- 深度学习和传统方法如何结合？
  - 深度神经网络的特征中引入传统方法：例如机器翻译任务中，将词的词性，句法，命名实体信息加入到输入特征。
  - 深度神经网络的设计借鉴传统方法：神经网络机器翻译的Attention机制，就是借鉴传统方法中的词对齐。

谢谢！