**（一）自然语言处理中的深度学习：方法及应用**

**Deep Learning in NLP: Methods and Applications**

**--林奕欧、雷航、李晓瑜、吴佳**

**（电子科技大学学报,2017,46(06):913-919.）**

**一、科学问题**

**1.1 本文所涉及科学问题**

讨论深度学习在自然语言处理方面的应用，分析其中的应用难点和突破方向。讲述了词向量的发展，深度神经网络模型的种类和原理。

**1.2 同行专家如何解决**

2013年，词向量word2vec兴起，各种词的分布式特征相相关研究层出不穷；2014年，研究者使用不同的DNN模型在NLP中取得重大进展；2015年，深度学习称为NLP主流工具。

**1.3 本文所解决的问题**

围绕特征表示和模型原理，以神经网络语言模型与词向量作为深度学习与自然语言处理结合的切入点，概述了深度神经网路的相关应用和最新的成果。最后总结了深度学习方法在当前自然语言处理研究应用中所遇到的困难和提出了未来的发展方向。

**1.4 本文解决方案效果**

在机器翻译领域，将DNN和词编码用于机器翻译，困惑度下降15%；利用双语词向量作为特征，对应的BLEU值提升了0.48；只使用字符序列训练循环神经网络文本生成器，效果接近加入了大量人工规则的文本生成系统。对于语义角色标注任务，使用半监督学习训练词编码，有效地提升了准确率。

**二、研究内容**

**2.1 理论与方法介绍**

深度信念网络( DBN)是由受限玻尔兹曼机(RBM)堆叠而生成的一种型。DBN通过训练网络的权重，使网络具有还原输入层训练数据的能力。DBN采用的训练步骤如下：

1) 当前层RBM为可见层则接收原始数据输入，否则接收上一层RBM的输出，并训练当前层RBM；2) 网络总层数满足要求则执行步骤4)，否则置下一层RBM为当前层；3) 重复步骤1)和步骤2)；4) 微调网络，使用有监督学习算法将模型收敛到局部最优解。

循环神经网络( RNN)相较于前馈神经网络，RNN可将本次隐藏层的计算结果用于下次隐藏层的计算。卷积神经网络( CNN)具有同一种叫池化的技术，在固定大小的区域使用平均值或最大值代替原有的矩阵区域。

**2.2 验证分析与实验效果**

循环神经网络的优化算法为BPTT算法。在循环神经网络的基础上提出长短时记忆模型(LSTM)。LSTM使用Cell结构记忆之前的输入，使得网络可以学习到合适的时机重置Cell结构，使用双向LSTM网络结合卷积神经网络和条件随机场解决词性标注和命名实体识别问题，分别取得97.55%91.21%的最好结果。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

除了仿生学的角度，目前深度学习的理论依据还处于起步阶段，大部分的研究成果都是经验性的，没有足够的理论来指导实验，研究者无法确定网络架构、超参数设置是否已是最优的组合。除此之外，目前仍没一种通用的深度网络或学习策略可以适用于大多数的应用任务。

**3.2 后续研究重点**

1) 可广泛适用于不同NLP任务的通用语义特征。

2) 超参数设置相关研究。

3) 新型网络架构和学习策略的提出和研究。

4)基于自然语言的逻辑推理和多模态应用。前者将提升机器的“智能”，后者扩展“智能”的应用领域。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：深度学习的昨天、今天和明天

**（2）作者介绍**：余凯、贾磊、陈雨强、徐伟

**（3）摘要**: 机器学习是人工智能领域的一个重要学科。自从20世纪80年代以来,机器学习在算法、理论和应用等方面都获得巨大成功。2006年以来,机器学习领域中一个叫"深度学习"的课题开始受到学术界广泛关注,到今天已经成为互联网大数据和人工智能的一个热潮。深度学习通过建立类似于人脑的分层模型结构,对输入数据逐级提取从底层到高层的特征,从而能很好地建立从底层信号到高层语义的映射关系。近年来,谷歌、微软、IBM、百度等拥有大数据的高科技公司相继投入大量资源进行深度学习技术研发,在语音、图像、自然语言、在线广告等领域取得显著进展。从对实际应用的贡献来说,深度学习可能是机器学习领域最近这十年来最成功的研究方向。将对深度学习发展的过去和现在做一个全景式的介绍,并讨论深度学习所面临的挑战,以及将来的可能方向。

**4.2 相关论文二**

**（1）题目**：**Extracting deep bottleneck features using stacked auto-encoders**

**（2）作者介绍**：[Gehring, Jonas](http://scholar.cnki.net/result.aspx?q=%e4%bd%9c%e8%80%85:(Gehring,+Jonas)&uid=WEEvREcwSlJHSldRa1FhdkJkVG1BKy9YMWgrOXJoYXVNNTRlTUhaUXdYOD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4ggI8Fm4gTkoUKaID8j8gFw!!&UID=WEEvREcwSlJHSldRa1FhdkJkVG1BKy9YMWgrOXJoYXVNNTRlTUhaUXdYOD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4ggI8Fm4gTkoUKaID8j8gFw!!" \t "http://d.scholar.cnki.net/detail/_blank)， [Miao, Yajie](http://scholar.cnki.net/result.aspx?q=%e4%bd%9c%e8%80%85:(Miao,+Yajie)&uid=WEEvREcwSlJHSldRa1FhdkJkVG1BKy9YMWgrOXJoYXVNNTRlTUhaUXdYOD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4ggI8Fm4gTkoUKaID8j8gFw!!&UID=WEEvREcwSlJHSldRa1FhdkJkVG1BKy9YMWgrOXJoYXVNNTRlTUhaUXdYOD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4ggI8Fm4gTkoUKaID8j8gFw!!" \t "http://d.scholar.cnki.net/detail/_blank) ,  [Metze, Florian](http://scholar.cnki.net/result.aspx?q=%e4%bd%9c%e8%80%85:(Metze,+Florian)&uid=WEEvREcwSlJHSldRa1FhdkJkVG1BKy9YMWgrOXJoYXVNNTRlTUhaUXdYOD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4ggI8Fm4gTkoUKaID8j8gFw!!&UID=WEEvREcwSlJHSldRa1FhdkJkVG1BKy9YMWgrOXJoYXVNNTRlTUhaUXdYOD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4ggI8Fm4gTkoUKaID8j8gFw!!" \t "http://d.scholar.cnki.net/detail/_blank) ，[Waibel, Alex](http://scholar.cnki.net/result.aspx?q=%e4%bd%9c%e8%80%85:(Waibel,+Alex)&uid=WEEvREcwSlJHSldRa1FhdkJkVG1BKy9YMWgrOXJoYXVNNTRlTUhaUXdYOD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4ggI8Fm4gTkoUKaID8j8gFw!!&UID=WEEvREcwSlJHSldRa1FhdkJkVG1BKy9YMWgrOXJoYXVNNTRlTUhaUXdYOD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4ggI8Fm4gTkoUKaID8j8gFw!!" \t "http://d.scholar.cnki.net/detail/_blank)

**（3）摘要**：在这项工作中，提出了一种用于从深度神经网络生成瓶颈特征的新型训练方案。 一堆去噪自动编码器首先以分层，无监督的方式训练。 然后，添加瓶颈层和附加层，并对整个网络进行微调以预测目标音素状态。 我们在广东话对话电话语音库中进行实验，发现增加网络中自动编码器的数量会产生更多有用的功能，但需要预训练，特别是在可用的训练数据很少的情况下。 使用更多未标记的数据进行预训练只会产生额外的收益。 对较大数据集和不同系统设置的评估证明了我们方法的普遍适用性。 在单词错误率方面，相对改进了9.2％（广东话，ML培训），9.3％（塔加路语，BMMI-SAT培训），12％（塔加拉族语，与MFCC的混淆网络组合）和8.7％（交换机）。

**4.3 相关论文三**

**（1）题目：Natural language neural network and its application to question-answering system**

**（2）作者介绍**：Tsukasa,Sagara; Masafumi,Hagiwara;

**（3）摘要**：本文提出了一种新的神经网络来处理自然语言。大多数传统神经网络只能处理由几个单词组成的句子，并且它们的应用非常简单，例如隐喻理解。所提出的网络可以处理许多复杂的句子，并且可以用作关联存储器和用于事实问题的问答系统。所提出的网络由三层和一个网络组成：句子层，知识层，深层案例层和字典网络。输入句子被分成知识单元并存储在知识层中。深层案例层在正确处理知识单元方面发挥着重要作用。字典网络作为知识库也发挥着重要作用。我们已经进行了几次实验，他们已经证明所提出的神经网络作为联想记忆和问答系统具有优越的性能。特别是作为一个问答系统，其性能非常接近基于人工智能的精心设计的系统。

**（二）用自然语言处理改进问答系统的研究和实现**

**Study and Implementation of Improving QA Systems Using NLP**

**--陈哲、文敦伟**

**（计算机工程,2006(20):205-206+223.）**

**一、科学问题**

**1.1 本文所涉及科学问题**

基于HMM模型，句法分析多种自然语言处理算法，将这些算法应用到问答系统中，改善系统的处理能力。

**1.2 同行专家如何解决**

目前许多的问答系统都使用了模式匹配的方法来进行语义理解，比如著名的ALICE 聊天机器人。

在一些使用问答系统作为用户查询接口的系统中，简单地将用户的输入作为系统询问的查询条件，同样具有局限性。

**1.3 本文所解决的问题**

基于模式匹配的问答系统不适应句子的变化，对于复杂句子结构的分析能力低下。为了适应更复杂的变化，需要对用户的语言进行语法和语义分析以及数据提取。

**1.4 本文解决方案效果**

将算法应用到问答系统中，扩展了对话管理模块，对用户的自然语言进行语义分析，从而实现对话的语义块识别、主题识别和对话信息提取，提高了系统对复杂用户输入的处理能力，克服了传统方法的不足，并使用Java 实现了一个实验系统，改善了传统模式匹配方法和将规则编入程序方法的弱点。

**二、研究内容**

**2.1 理论与方法介绍**

本文研究了基于词典和规则的语法分析算法，使得系统具有更强的适应性。系统包括词性标注算法、语法分析算法。

自然语言句子中的词性和单词的位置有很密切的联系，如形容词经常出现在名词前面。因此用一个马尔可夫模型(HMM)来描述词性概率。

句法分析算法使用改进的图句法分析算法。有n 个词语的语句的图由n＋1 个顶点和多条连接这些顶点的边组成，边如果与它相邻的边匹配，就将边扩展后插入图中，这样，算法结束时，只要得到从第1 个节点到最后一个节点的边（称为顶点边），就可以得到一棵语法分析树，通过这棵树就可以遍历所有的边。

**2.2 验证分析与实验效果**

基于槽的问答系统有几种对回答的理解方式：(1)用单个的名词来回答。(2)采用“提问点＋回答”的双名词形式。(3)一句话中包含多个语义块，而每个语义块都包含了对一个槽的信息。

针对（1）只需要将用户的输入作为值填充到当前提问的槽中即可。

针对（2）需要同时识别句子中的槽和值。首先将句法分析树中的NP全部识别出来，然后进行槽和值的识别，就可以得出结果。

针对（3）将复杂句子分成若干个单语句，在后面的步骤中分别处理。将该算法中得到的若干个语义块根边节点分别使用槽和值识别算法就完成了对用户多语义块自然语句的识别和数据信息提取。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

系统的字典和词性库不会自己进行维护，对用户的语言习惯没有过多地进行训练

**3.2 后续研究重点**

(1)对算法的改进，考虑使用机器学习的方法对系统进行扩展，使得系统可以自己维护字典和词性概率矩阵，包括新词学习等。(2)进一步对用户语言习惯进行研究和分析，完善自然语言处理算法。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：基于深度学习的智能问答系统研究与设计

**（2）作者介绍**：赵昌志、王怡婷、张小琴

**（3）摘要**: 智能客户服务系统是在大规模知识处理的基础上开发的面向行业的应用程序,适用于大规模知识处理、自然语言理解、知识管理、自动问答系统和推理等技术行业。近年来,机器学习、强化学习、深度学习等技术的革命性发展,促进了问答系统的进一步发展。笔者设计了一个系统,该系统基于给定文档提取知识,并自动生成知识库,实现基本功能。

**4.2 相关论文二**

**（1）题目**：基于自然语言处理的医疗健康问答系统

**（2）作者介绍**：颜昕

**（3）摘要**：随着医疗行业用人紧张,医患矛盾的社会现象出现,为了让患者(或潜在患者)及时了解到自己的身体健康情况,并且得到专业的、人性化的诊断回答。本文基于这一社会背景,结合人工智能的火热发展趋势,提出了使用基于自然语言处理的方法,并运用多种机器学习的模型训练海量的数据,还原真实医患场景。最终拟设计生成一款关于医疗健康的智能问答系统方案,为居家医疗等模式提供最大的便利。

**4.3 相关论文三**

**（1）题目：**基于自然语言处理的医院智能客服系统的设计与研究

**（2）作者介绍**：李楠

**（3）摘要**：目的:利用机器学习和自然语言处理等技术,实现医院客服系统从模板式应答向智能学习式转型。方法:对市面上常见的客服系统进行归纳分类,原理阐述、优劣辨析、需求汇总并理出框架。在此基础上,借助"图灵机器人"平台搭建智能客服系统环境,并完成测试。结果:基于机器学习、语义分析等技术为医院量身定制的智能客服系统,可实现全天候、全自动为患者提供精准的答案和智能化的服务。结论:医院作为一个专业性强、垂直分科多、人流密度大的公共服务场所,引入无人值守的智能客服系统,既能显著节约人工客服的工作量,更能满足业务需求,提升患者体验。