北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.011

基于 Deep Learning 的代词指代消解

奚雪峰[1,2]，　周国栋[1,†]

1. 苏州大学计算机科学与技术学院　自然语言处理实验室，苏州　215006; 2. 苏州科技学院　计算机科学与工程系，苏州　215009; † 通信作者, E-mail: gdzhou@suda.edu.cn

摘要　指代消解一直是自然语言处理中的核心问题，提出一种利用 DBN (deep belief nets)模型的 Deep learning 学习机制进行基于语义特征的指代消解方法。DBN 模型由多层无监督的 RBM (restricted Boltzmann machine)网络和一层有监督的 BP (back-propagation)网络组成, RBM 网络确保特征向量映射达到最优，最后一层 BP 网络分类 RBM 网络的输出特征向量，从而训练指代消解分类器。在 ACE04 英文语料及 ACE05 中文语料上进行测试，实验结果表明，增加 RBM 训练层数可以提高系统性能。此外，引入对特征集合的抽象分层因素，也对系统性能提升产生积极作用。

关键词　代词消解；深度学习；深层语义特征

中图分类号　TP391

Pronoun Resolution Based on Deep Learning

XI Xuefeng[1,2], ZHOU Guodong[1,†]

1. Natural Language Processing Laboratory, School of Computer Science and Technology, Soochow University, Suzhou 215006;
2. Department of Computer Science and Engineering, Suzhou University of Science and technology, Suzhou 215009;
† Corresponding author: E-mail: gdzhou@suda.edu.cn

Abstract　Because coreference resolution is a fundamental task in natural language process, a coreference resolution system based on deep learning model via the deep belief nets (DBN), which is a classifier of a combination of several unsupervised learning networks, named RBM (restricted Boltzmann machine) and a supervised learning network named BP (back-propagation), is proposed to detect and classify the coreference relationships between the anaphor and antecedent. The RBM layers maintain as much information as possible when feature vectors are transferred to next layer. The BP layer is trained to classify the features generated by the last RBM layer. The experiments are conducted on the ACE 2004 English NWIRE corpus and the ACE 2005 Chinese NWIRE corpus. The results show that increasing the number of layers RBM training and joining of abstract layer for feature set are able to improve the performance of coreference resolution system.

Key words　pronoun resolution; Deep Learning; deep semantic feature

自然语言篇章中的一个语言单位(在语言学中称为照应语 anaphor)与之前出现的语言单位(称为先行语 antecedent)之间所存在的特殊语义关联，称为指代，确定照应语所指的先行语的过程即为指代消解。根据先行语的不同类型，指代可以分为实体指代和事件指代。如图 1 例子所示。例子句中的“他”指代“本·拉登”，属于一类实体指代；而例子句中的“这”指代“炸毁”这一事件，属于事件指代。要理解篇章的整体含义，首先需要明确句子中照应语(如代词“他”、“这”)与前句先行语的指代关系。

因此，指代关系实质上搭建起篇章的话题结构，从内容层面实现篇章内句与句之间的衔接性(cohesion)。准确无歧义的指代消解保证对篇章语义的整体理解，这对信息抽取(information extrac-

国家自然科学基金(61273320, 61003153, 61272257)和 863 计划（2012AA011102)资助
收稿日期: 2013–06–24; 修回日期: 2013–09–16; 网络出版时间: 2013-11-11 10:25

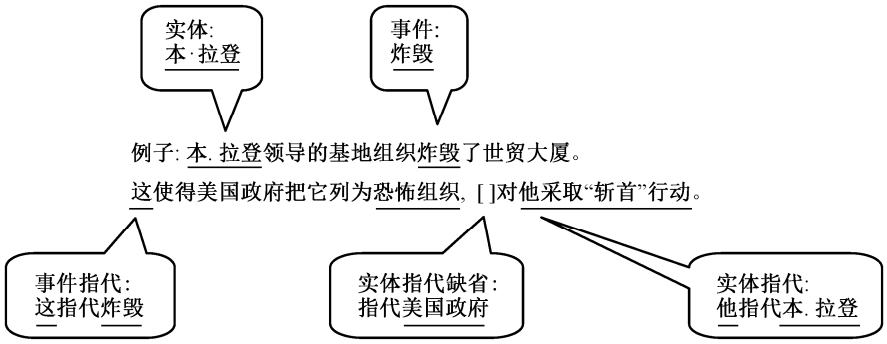

图 1　篇章级指代消解的例子

Fig. 1　An example of coreference resolution in a discourse

tion)、自动文摘(automatic summarization)、自动问答(question answering)以及机器翻译(machine translation)等自然语言应用系统都有极为重要的基础支撑作用。

指代消解的研究方法大致可以分成采用规则的方法、采用统计的方法和采用分类的方法。在使用语料方面, 早期通过手工构建领域知识, 形成消解规则进行指代消解。近十年来, 随着标注语料库的不断出现以及 Internet 的迅速发展, 实验语料的获得越来越方便, 目前大多数的指代消解研究已转向基于语料库的指代消解方法。在指代消解中采用分类方法, 始于 McCarthy 等[1]把判断先行语的问题转换成分类问题, 通过分类器判断指代语与每个先行语候选之间是否存在指代关系。Soon 等[2]借鉴这一思想, 基于机器学习方法提出一个完整的指代消解框架。在此基础上, Ng 等[3]对 Soon 的研究进行扩充, 探索一系列词法、语法和语义特征, 所构建的系统取得显著性能。Zhou 等[4]对先行语候选指代链中的语义信息在代词(特别是中性代词)指代消解中所起的作用进行探索, 并在此基础上提出进一步使用上下文信息和网络挖掘技术自动判别代词的语义类别的方法, 更好地解决了代词的指代消解。Ng[5]详细探讨各种语义信息对指代消解的意义。Kong 等[6]以语义角色为载体, 将中心理论从传统的语法层拓展到语义层, 进一步探索语义信息对代词消解性能的作用。研究表明, 有效的语义信息能够极大地提升英文指代消解的性能。与英文相比, 目前中文指代消解的研究相对较少。王厚峰等[7]根据中文人称代词的语义角色和对应的先行语可能的语义角色, 给出消解人称代词的基本规则; 采用近似 Mitkov 的基于弱化语言知识的方法来解决人称代词的消解[8]。许敏等[9]利用格框架, 提出在上下文相关语义环境中进行指代分类解决的思想, 并给出相应的算法。王海东等[10]探索语义角色对指代消解性能的影响; 孔芳等[11]利用卷积树核函数, 探索语义信息对中文代词消解性能的作用。研究表明, 语义角色信息的引入同样也能够显著提高中文指代消解的性能。

随着指代消解研究的不断深入, 越来越多的研究者发现, 语义信息在指代消解中起至关重要的作用。而在语义信息处理中, 传统常见的解决方法就是使用一定的模型来表示语义信息, 并依靠人工经验来抽取样本特征规则, 基于机器学习方法实现分类或预测。虽然这种方法在一定程度上提升了系统的性能, 但是存在以下局限性。

1) 在模型运用不出差错的前提下, 特征规则的好坏就成为整个系统性能的瓶颈。人工设计样本特征规则时, 若要发现一个好的特征规则, 就要求研究人员对待解决的问题有很深入的理解。而达到这个程度, 往往需要反复摸索, 耗时太久。因此, 人工设计样本特征规则, 存在可扩充性差的局限。

2) 上述采用的传统机器学习, 大多属于一类浅层学习方法, 如隐马尔科夫模型(HMM)、条件随机场(CRFs)、最大熵模型(ME)、支持向量机(SVM)、核回归及仅含单隐层的多层感知器等, 这些模型的结构基本上可以看成带有一层隐层节点(如 SVM), 或没有隐层节点(如 ME)。仅含单层非线性变换的浅层学习结构, 其局限性在于有限样本和计算单元情况下对复杂函数的表示能力有限, 面临复杂问题处理时的泛化能力受到一定制约[12]。

近年来在机器学习领域出现的深度学习(Deep Learning)方法[13], 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, 展现从少数样本集中学习数据集本质特征的强大能力。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。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 3 层、5 层, 甚至 10 多层的隐层节点; 2) 明确突出特征学习的重要性, 通过逐层特征变换, 将样本在原空间的特征表示变换到一个新特征空间, 使分类或预测更加容易。

基于上述原因, 本文提出利用 Deep Learning 方法进行基于深层语义的指代消解。研究 Deep Learning 深层学习结构的处理机制, 探索面向指代消解的语义特征泛化表示, 利用 Deep Learning 深层学习机制自动挖掘深层语义信息, 研究深层语义信息在指代消解中的作用。由于代词在指代消解中占有重要地位, 本文将重点研究代词的指代消解。

1 相关工作

在处理大量感知数据的过程中, 人类总能够高效准确地获取到值得注意的重要信息。例如, 即使是 3 岁幼童, 下班时间站在小区门口观望繁杂的回家人群, 也总是能够快速准确地发现妈妈熟悉的身影。模仿人脑那样高效准确地处理信息一直是人工智能研究领域的核心挑战。基于哺乳类动物大脑的解剖学知识, 神经科学研究人员通过测试感官信号从视网膜传递到前额大脑皮质再到运动神经的时间, 推断出大脑皮质并未直接地对数据进行特征提取处理, 而是使接收到的刺激信号通过一个复杂的层状网络模型, 进而获取观测数据展现的规则[14−16]。也就是说, 人脑并不是直接根据外部世界在视网膜上投影, 而是根据经聚集和分解过程处理后的信息来识别物体。因此视皮层的功能是对感知信号进行特征提取和计算, 而不仅仅是简单地重现视网膜的图像[17]。人类感知系统这种明确的层次结构极大地降低了视觉系统处理的数据量, 并保留了物体有用的结构信息。对于要提取具有潜在复杂结构规则的自然语言、图像、视频、语音和音乐等结构丰富数据, 深度学习能够获取其本质特征。

受大脑结构分层次启发, 神经网络研究人员一直致力于多层神经网络的研究。BP 算法是经典的梯度下降并采用随机选定初始值的多层网络训练算法, 但因输入与输出之间的非线性映射使得网络误差函数或能量函数空间是一个含多个极小点的非线性空间, 搜索方向仅是使网络误差或能量减小的方向, 因而经常收敛到局部最小, 并随网络层数增加情况更加严重。理论和实验表明 BP 算法不适于训练具有多隐层单元的深度结构[18]。此原因在一定程度上阻碍了深度学习的发展, 并将大多数机器学习和信号处理研究从神经网络转移到相对较容易训练的浅层学习结构。

深层学习结构研究的突破性进展是由 Hinton[13] 在 2006 年取得的, 几乎是在同一年, Bengio 等[19]和 Ranzato 等[20]迅速跟进, 从此开启 Deep Learning 的研究热潮。Deep Learning 的主要思想是把学习结构(learning hierarchy)看做是一个网络(network), 那么: 1) 无监督学习用于每一层网络的 pre-train; 2) 每次用无监督学习只训练一层, 并将其训练结果作为其更高(更抽象)一层的输入; 3) 用有监督学习去调整所有层。最终, 构建一个深度有监督学习的分类器, 如神经网络分类器, 或者构建一个深度生成模型, 如 DBM (Deep Boltzmann Machine) [21]。

深度学习具有多层非线性映射的深层结构, 其优势之一是可以完成复杂的函数逼近; 此外深度学习理论上可获取分布式表示, 即可通过逐层学习算法获取输入数据的主要驱动变量[22]。该优势是通过深度学习的非监督预训练算法完成, 通过生成性训练可避免因网络函数表达能力过强而出现过拟合情况。但是单层计算能力有限, 因此, 深度学习通过多层映射单元提取出主要的结构信息。

深度学习目前在自然语言处理的诸多领域, 如情感分析、句法分析、文本蕴涵、实体关系抽取等, 有着初步应用。Glorot 等[23]分析了目前指数级增长的网络推荐、评论等数据信息, 认为这些数据信息覆盖领域过于广泛, 很难获取针对某个领域的、有意义的标注数据。在进一步研究领域自适应的情感分类器的基础上, Glorot 等提出一种深度学习方法, 该方法采用一种无监督学习方式, 从网络评论数据中学习如何提取有意义的信息表示, 并将其应用于构建一个情感分类器系统, 在 Amazon 产品的 4 类评论基准数据测试上取得了显著性能。Collobert 等[24] 提出一种通用的深层神经网络结构和学习算法, 此算法能够应用于各类自然语言处理任务, 如词性标注、chunking、命名实体识别和语义角色标注。这种算法之所以具备多功能通用性, 原因在于忽略了具体任务的特殊性以及众多领域知识。不同于常见的通过人工提取优化的特征值的方式, 该算法基于各类未标注训练数据来学习内在表示信息。在某个标注系统中应用该算法, 性能结果及计算速度上都

有上佳表现。陈宇等[25]提出一种利用 DBN (Deep Belief Nets)模型进行基于特征的实体关系抽取方法，实验结果表明，DBN 非常适用于基于高维空间特征的信息抽取任务。

尽管深度学习在上述自然语言处理中有了初步应用，但是在指代消解领域，还未见到有利用该方法来抽取深层语义信息从而应用于指代消解的文献报道。本文利用 Deep Learning 深层学习机制自动挖掘深层语义信息，研究深层语义信息在指代消解中的作用。

2 基于 Deep Learning 的代词消解

本节从系统框架、特征表示和样例生成这 3 方面入手，介绍基于 Deep Learning 的中英文代词消解方案，并通过对 ACE 2004 NWIRE 英文语料和 ACE 2005 NWIRE 中文语料上实验结果的分析，探索深层语义特征信息对代词消解的作用。

2.1 系统框架

针对英文及中文的指代消解问题，本文基于统一的系统框架，采用全自动的方式实现指代消解。参考 Soon 等[2]提出的指代消解基本框架结构，构成如图 2 所示具体框架。由于中英文语言体系的差异，在某些模块的处理上稍有不同。

在英文平台上，我们使用基于隐马尔可夫模型的命名实体识别、词性标注和名词短语识别模块[26−27]对语料进行预处理。在中文指代消解平台上，使用 Stanford(http://nlp.stanford.edu)的中文分词和词性标注模块以及基于可信模型的命名实体识别模块(苏州大学自然语言处理实验室，http//nlp.suda.edu.cn)对语料进行线性预处理。训练和测试中，首先采用与文献[11]一致的方式进行实例的初步生成，之后引入一层特征抽象值。

2.2 特征表示

众多研究成果已经表明，语义特征信息对指代消解，特别是代词消解具有重要意义，但哪些语义特征信息对代词消解是直接有效的，这一问题依然没有很好解决。传统浅层机器学习方法结合特征规则进行处理，强调人工提取有效特征规则，在方法不出错的情况下，面向领域问题的特征规则就变得非常关键。然而从人类认知过程的研究来看，上述传统方法至少存在特征规则可扩充性差及浅层学习方法泛化表示能力弱的两个主要问题。Deep Learning 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，自动学习更有用的特征，从而提升分类或预测的准确性。

Soon 等[2]首次抽取出 12 个表层特征，并基于机器学习方法提出一个完整的指代消解框架。在此基础上，Ng 等[3]对 Soon 的研究进行了扩充，共抽取出 53 个系列词法、语法和语义特征，在国际标准评测系统 MUC-6 上取得 F 值 69.4 的显著性能。Yang 等[28]探索候选语指代属性在消解中的作用，抽取出 24 个先行语和候选语的词法、句法、语义及位置特征，取得优于 Baseline 系统的消解性能。上述系统中抽取出的特征表示，尽管都考虑不同层面的特征，如词法/语法/语义/位置层面等，但是在利用特征的实际学习过程中，依然采用单层学习机制，仅考虑特征值的作用，而没有考虑特征层次类型在系统学习中的影响。事实上，人类认知过程带有抽象分层特性，对于某些未知物体的识别，往往是从抽象到具体，不同层次归类。抽象层次越高，物体识别的准确率越高。因此，有必要对特征类型进行分层，并发挥其在学习过程中的作用。

本文在英文指代消解平台上，采用文献[28]的 24 个特征集，但是扩充定义特征抽象层，如表 1 所示。根据特征所属不同抽象层次，分别定义 5 类不同特征抽象层次，赋予不同的特征抽象值，并将其引入 Deep Learning 的训练数据集中，进一步促进 Deep Learning 抽象分层学习性能。英文平台使用特征集定义如表 2 所示。中文平台使用文献[11]所定义的 17 个特征集，如表 3 所示，同样采用上述定义的特征抽象值进行扩充。

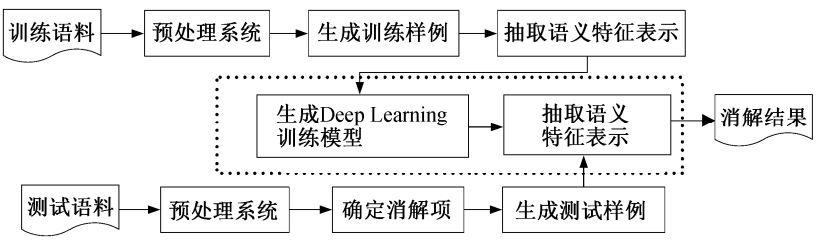

图 2　系统框架

Fig. 2　Framework of propesod system

表 1　特征抽象层

Table 1　Feature levels

特征抽象层	抽象值
Lexical	1
Grammatical	2
Semantic	3
Position	4
Other	0

2.3　样例生成

本文采用 Soon 等[2]的方法, 首先通过聚类形成链, 之后采用分类方法实现分类。例句“我曾直接要求藤森逮捕西蒙, 并立即把他送上法庭受审。”经预处理后形成的中文训练及测试样例格式例子如表 4 所示。其中中文平台 17 个特征值对应特征依次为: ANPronoun, ANDefiniteNP, ANDemonstrativeNP, ANPronounType, CAPronoun, CAPronounType, CAARG0, CAARG0MainVerb, ANCABothProperName, ANCANameAlias, ANCASentDistance, ANCAGenderAgreement, ANCANumberAgreement, ANCAAppositive, ANCAHeadStringMatch, ANCAWORDSENSE, ANCASameTarget。另外, 抽象值可查表 3 对应得到。英文平台与中文平台类同。

表 2　英文指代消解平台的特征集

Table 2　Feature set of the English platform in coreference resolution

抽象值	特征抽象层	特征	描述对象	特征值定义
1	Lexical	candi_DefNP	先行语候选词	若先行语候选词是个有定名词短语, 则取 1; 否则取 0
1	Lexical	candi_DemoNP	先行语候选词	若先行语候选词是个非定名词短语, 则取 1; 否则取 0
1	Lexical	candi_Pron	先行语候选词	若先行语候选词是个代名词, 则取 1; 否则取 0
1	Lexical	candi_ProperNP	先行语候选词	若先行语候选词是个固有名字, 则取 1; 否则取 0
1	Lexical	candi_NE_Type	先行语候选词	若先行语候选词是个组织机构命名实体, 则取 1; 如果是人名, 则取 2; 其他命名实体, 则取 3; 不是命名实体, 则取 0
3	Semantic	candi_Human	先行语候选词	根据 WordNet 抽取出, 候选词是一个人名实体的概率, 在 0-100 之间取值
2	Grammatical	candi_FirstNPInSent	先行语候选词	若候选词是所在句子中的第一个名词短语, 则取 1; 否则取 0
4	Position	candi_Nearest	先行语候选词	若候选词距离照应语最近, 则取 1; 否则取 0
2	Grammatical	candi_SubjNP	先行语候选词	若候选词是当前句子的主语, 则取 1; 否则取 0
1	Lexical	ana_Reflexive	照应语	若照应语是一个反身代名词, 则取 1; 否则取 0
2	Grammatical	ana_Type	照应语	若照应语是一个第三人称代名词(he, she, ...), 则取 1; 若是单数中性代名词(it, ...)则取 2; 若是复数中性代名词(they, ...), 则取 3; 其他类型取 4
4	Position	SentDist	两者关系	表示照应语和先行语在一个句子内的距离值
4	Position	ParaDist	两者关系	表示照应语和先行语在一个段落内的距离值
2	Grammatical	CollPattern	两者关系	若先行语与照应语词序排列模式相同, 则取 1; 否则取 0
1	Lexical	ante-candi_DefNp	候选词的先行语	若候选词的先行语是一个有定名词短语, 则取 1; 否则取 0
1	Lexical	ante-candi_InDefNp	候选词的先行语	若候选词的先行语是一个非定名词短语, 则取 1; 否则取 0
1	Lexical	ante-candi_Pron	候选词的先行语	若候选词的先行语是一个代名词, 则取 1; 否则取 0
1	Lexical	ante-candi_Proper	候选词的先行语	若候选词的先行语是一个固有名称, 则取 1; 否则取 0
1	Lexical	ante-candi_NE_Type	候选词的先行语	若候选词的先行语是个组织机构命名实体, 则取 1; 如果是人名, 则取 2; 其他命名实体, 则取 3; 不是命名实体, 则取 0
3	Semantic	ante-candi_Human	候选词的先行语	根据 WordNet 抽取出, 候选词的先行语是一个人名实体的概率, 在 0~100 之间取值
2	Grammatical	ante-candi_FirstNPInSent	候选词的先行语	若候选词的先行语是所在句子中的第一个名词短语, 则取 1;否则取 0
2	Grammatical	ante-candi_SubjNP	候选词的先行语	若候选词的先行语是句子的主语, 则取 1; 否则取 0
2	Grammatical	Apposition	两者关系	若候选词的先行语与候选词属于同位格结构, 则取 1; 否则取 0
3	Semantic	candi_NoAntecedent	候选词	若候选词没有可对应的先行语, 则取 1; 否则取 0

3　研究模型

目前基于 Deep learning 理论的应用系统中, 深度可信网络(Deep Belief Nets, DBN)是比较广泛的

一类学习结构, 它由多层受限波尔兹曼机(RBM)单元和一层有监督网络层组成。本文采用 DBN 来构建面向指代消解的深度学习计算平台。

3.1 深度可信网络(DBN)训练

本文所采用的 DBN 是由若干层自底向上的 RBM 和一层有监督的 BP(Back-Propagation)网络组成的深层神经网络[13], 其结构如图 3 所示[25]。$\boldsymbol{v}_i$ 和 $\boldsymbol{h}$ 分别表示可视层和隐含层内的节点值, W 表示可视层和隐含层之间的权值。底层的神经网络接收原始的特征向量, 在自底向上的传递过程中, 从具体的特征向量逐渐转化为抽象的特征向量, 在顶层的神经网络形成更易于分类的组合特征向量。增加网络层数能够将特征向量更加抽象化[29]。而且, 尽管 RBM 确保训练后的层内参数达到最优, 但却不能完全消除映射过程中产生的错误和不重要的信息, 多层神经网络的每一层网络会弱化上一层网络产生的错误信息和次要信息, 因此, 深层网络较单层网络精确度更高。

在训练模型的过程中, DBN 主要分为两步: 首先分别单独无监督地训练每一层 RBM 网络, 然后在最后一层设置有监督 BP 网络分类器。第一步的作用在于确保特征向量映射到不同特征空间时, 都尽可能多地保留特征信息; 第二步的作用在于通过 BP 网络接收 RBM 的输出特征向量作为它的输入特征向量, 有监督地训练分类器。 此外, 为确保每一层的联合概率分布 $p(v, h)$最大, 必须使当前层 RBM 网络调整自身层内的权值对该层特征向量映射达到最优; 但显然仅依靠 RBM 层并不能对整个 DBN 的特征向量映射达到最优。所以, BP 网络还担负着微调功能: BP 网络将错误信息自顶向下传播至每一层 RBM, 可以达到微调整个 DBN 网络的作用。另一方面, DBN 引入 RBM 网络训练模型的过程, 实现了对一个深层 BP 网络权值参数的初始化, 克服了 BP 网络因随机初始化权值参数而容易

表 3 中文指代消解平台的特征集

Table 3 Feature set of the Chinese platform in coreference resolution

抽象值	特征抽象层	特征	描述对象	特征值定义
1	Lexical	ANPronoun	照应语	若照应语是代词, 则取 1; 否则取 0
1	Lexical	ANDefiniteNP	照应语	若照应语是有定名词短语, 则取 1; 否则取 0
1	Lexical	ANDemonstrativeNP	照应语	若照应语是指示性名词短语, 则取 1; 否则取 0
1	Lexical	ANPronounType	照应语	照应语若为代词, 其具体的代词细分类别
1	Lexical	CAPronoun	先行语	若先行语是代词, 则取 1; 否则取 0
1	Lexical	CAPronounType	先行语	先行语若为代词, 其具体的代词细分类别
3	Semantic	CAARG0	先行语	若先行语承担 Arg0 语义角色, 则取 1; 否则取 0
3	Semantic	CAARG0MainVerb	先行语	若先行语承担的 Arg0 语义角色是由主谓词驱动, 则取 1; 否则取 0
1	Lexical	ANCABothProperName	两者关系	若照应语和先行语候选词均为专有名词, 则取 1; 否则取 0
1	Lexical	ANCANameAlias	两者关系	若照应语和先行语候选词存在别名关系, 则取 1; 否则取 0
4	Position	ANCASentDistance	两者关系	若照应语和先行语在 1 句内取 1, 2 句取 0.9, …, 大于 10 句取 0
2	Grammatical	ANCAGenderAgreement	两者关系	若照应语和先行语满足词性一致, 则取 1; 否则取 0
2	Grammatical	ANCANumberAgreement	两者关系	若照应语和先行语满足单复数一致, 则取 1; 否则取 0
2	Grammatical	ANCAAppositive	两者关系	若照应语和先行语是同位语, 则取 1; 否则取 0
2	Grammatical	ANCAHeadStringMatch	两者关系	若照应语和先行语满足中心词匹配, 则取 1; 否则取 0
3	Semantic	ANCAWORDSENSE	两者关系	若从 HowNet 中获得的语义信息类有相同的, 则为 1; 否则为 0
3	Semantic	ANCASameTarget	两者关系	若先行语和照应语承担的语义角色是由相同谓词驱动, 则取 1; 否则取 0

表 4 中文平台上的训练及测试样例格式

Table 4 Pairs of markable for training or testing in Chinese Platform

先行语	照应语	样例值(17 个特征值+17 个抽象值)	是否指代
西蒙	他	1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0.9, 1, 1, 0, 0, 0, 0+1, 1, 1, 1, 1, 1, 3, 3, 1, 1, 4, 2, 2, 2, 2, 3, 3	是
藤森	他	1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0.9, 1, 1, 0, 0, 0, 0+1, 1, 1, 1, 1, 1, 3, 3, 1, 1, 4, 2, 2, 2, 2, 3, 3	否

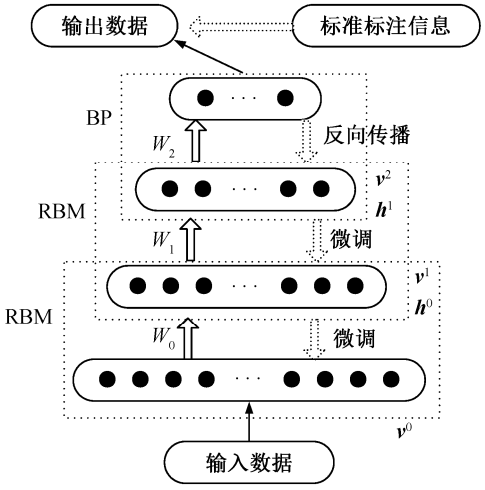

图 3　DBN 网络结构图
Fig. 3　Structure of DBN

陷入局部最优和训练时间长的缺点。DBN 训练过程实现步骤如算法 1 所示。

算法 1　DBN 训练过程算法

1) RBM 层数设定为 Rlevels, 并初始化当前 RBM 训练层 CL=1, 将输入特征向量赋值给 v^0;

2)将 v^0 作为 RBM 训练算法的输入向量, 调用算法 2(见 3.2 节算法 2)训练 RBM 权值等层数;

3)利用当前 RBM 网络输出转换后的特征向量值, 赋值给 v^0;

4)如果当前 RBM 训练层 CL=训练参数 RLevels, 则转步骤 5, 否则 CL=CL+1, 并转向步骤 2;

5)调用算法 3(见 3.3 节算法 3), 使用 BP 网络有监督地训练分类器, 并自顶向下反向微调整个 DBN 网络。

6)训练结束。

3.2　受限玻尔兹曼机(RBM)自训练

RBM 是 DBN 的核心组件之一, 它由一个可见层 V 和一个隐含层 H 组成, 层间的节点两两相连, 层内的节点不相连, 其结构如图 4 所示。隐单元可获取输入可视单元的高阶相关性。

RBM 训练中的关键问题是要获取生成性权值。由上定义, W 表示可视层和隐含层之间的权值, 令 $\boldsymbol{b}$ 和 $\boldsymbol{c}$ 分别表示可视层和隐含层的偏置量, 则需要确定参数 $\theta=(W, \boldsymbol{b}, \boldsymbol{c})$, 使得联合概率分布 $p(\boldsymbol{v}, \boldsymbol{h})$最大[30]。

利用式(1)可以由已知的可视层的节点值得到隐含层的节点值[25]:

$$p(h_j=1)=\frac{1}{1+\exp\left(-b_j-\sum_i v_i w_{ij}\right)}。\tag{1}$$

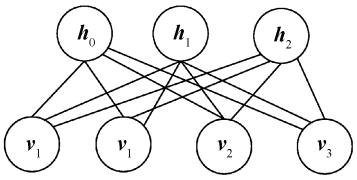

图 4　RBM 结构图
Fig. 4　Structure of RBM

由文献[13]可知, RBM 是对称网络, 同理利用式(2)可以由已知的隐含层的节点值得到可视层的节点值:

$$p(v_i=1)=\frac{1}{1+\exp\left(-c_i-\sum_j h_j w_{ji}\right)}。\tag{2}$$

此时, 隐含层内的特征向量 $\boldsymbol{h}$ 和可视层内的特征向量 $\boldsymbol{v}$ 的联合概率分布满足:

$$p(\boldsymbol{v},\boldsymbol{h})\propto\exp(-E(\boldsymbol{v},\boldsymbol{h}))=\mathrm{e}^{\boldsymbol{h}^{\mathrm{T}}W\boldsymbol{v}+\boldsymbol{b}^{\mathrm{T}}\boldsymbol{v}+\boldsymbol{c}^{\mathrm{T}}\boldsymbol{h}}\tag{3}$$

其中, $E(\boldsymbol{v}, \boldsymbol{h})$是特征向量 $\boldsymbol{v}$ 和特征向量 $\boldsymbol{h}$ 数学期望, 其绝对值的大小代表特征向量 $\boldsymbol{h}$ 保存着特征向量 $\boldsymbol{v}$ 的信息的多少。

为了求出使得联合概率 $p(\boldsymbol{v}, \boldsymbol{h})$最大的参数 θ, 传统的做法是利用马尔可夫链蒙特卡罗(Markov chain Monte Carlo, 简称 MCMC)。基本步骤是: 首先可视层和隐含层互为条件, 不断地求得更新状态, 最后共同趋向平稳状态, 此时联合概率 $p(\boldsymbol{v}, \boldsymbol{h})$达到最大值[31]; 之后求得最大联合概率分布与初始状态的联合概率分布的斜率; 最后用式(4)更新权值 θ。

$$\theta^{(\tau+1)}=\theta^{(\tau)}+\eta\left.\frac{\partial\log P(v,h)}{\partial\theta}\right|_{\theta^{\tau}},\tag{4}$$

其中, τ 为迭代次数, η 为学习速度。

RBM 的输入向量 $\boldsymbol{v}^0$ 是 t=0 时刻可视层的特征向量; $\boldsymbol{h}^0$ 是由 $\boldsymbol{v}^0$ 根据式(3)得到的隐含层特征向量; $\boldsymbol{v}^1$ 是 t=1 时刻可视层的特征向量, 由得到的 $\boldsymbol{h}^0$ 经式(2)计算得到。以此类推, $\boldsymbol{v}^\infty$和 $\boldsymbol{h}^\infty$分别是 t=∞时刻可视层和隐含层的特征向量。斜率可由式(5)计算得出:

$$\begin{aligned}\frac{\partial\log P(v,h)}{\partial\theta_{ij}}&=<h_j^0(v_i^0-v_i^1)>+<v_i^1(h_j^0-h_j^1)>+\ldots\\&=<h_j^0v_i^0>-<h_j^0v_i^1>+\\&\quad<v_i^1h_j^0>-<v_i^1h_j^1>+\ldots\\&=<h_j^0v_i^0>-<h_j^\infty v_i^\infty>,\end{aligned}\tag{5}$$

其中, $\boldsymbol{h}^0\boldsymbol{v}^0$ 为输入特征向量与其对应的隐含层特征向量的点乘的平均值; $\boldsymbol{h}^\infty\boldsymbol{v}^\infty$为马尔可夫链末端可视

层特征向量与其对应的隐含层特征向量的乘积的平均值, $\boldsymbol{h}^{\infty}\boldsymbol{v}^{\infty}$是收敛的。由式(5)可知, 联合概率分布的斜率与中间状态无关, 只与网络的初始状态和最终状态有关。根据式(4)可以得出修改后的参数 θ, 从而达到自训练的目的。

传统马尔可夫链的方法在求最佳联合概率 $p(\boldsymbol{v}^{\infty}, \boldsymbol{h}^{\infty})$和初始联合概率分布 $p(\boldsymbol{v}, \boldsymbol{h})$时, 收敛速度很难保证, 并且难以确定步长 ∞。Hinton[32]提出利用 Contrastive Divergence (CD)准则保持精度的同时能够快速提高计算速度。利用 Kullback-Leibler 距离衡量两个概率分布的"差异性", 表示为 KL(P||P'), 如式(6)所示:

$$\mathrm{CD}_n = \mathrm{KL}(p_0 \| p_\infty) - \mathrm{KL}(p_n \| p_\infty), \tag{6}$$

其中, p_0 为 RBM 网络初始状态的联合概率分布, p_n 为经过 n 步马尔可夫链之后的 RBM 网络的联合概率分布, p_∞为马尔可夫链末端的 RBM 网络的联合概率分布。CD_n 可以看做是 p_n 介于 p_0 和 p_∞之间的位置衡量。通过不断地将 p_n 赋值给 p_0, 得到新的 p_0 和 p_n。实验[33]证明, 在 r 次求斜率修正参数后, CD_n 趋向于 0, 且精度近似于马尔可夫链方法。因此, 本文采用基于 CD 准则 RBM 网络训练方法[25], 实现步骤如算法 2 所示。

算法 2 基于 CD 准则的 RBM 网络自训练过程

1) 随机初始化 θ_0=(W_0, b_0, c_0)赋值给 θ_t, 并设定迭代次数 Step;

2) 将输入特征向量赋值给 v^0, 并利用式(1)和(2)计算特征向量 $\boldsymbol{h}^0$, $\boldsymbol{v}^1$ 和 $\boldsymbol{h}^1$;

3) 利用式(5)得到 RBM 网络初始状态与更新状态下的联合概率分布的斜率, 并代入式(4)修正参数 θ_t, 得到 θ_{t+1};

4) 如果 t=Step, 程序结束; 如果 t<Step, 则将 θ_{t+1} 赋值于 θ_t, 并转步骤 2。

3.3 BP 网络训练

BP 网络是一种有监督分类器, 在 DBN 的最后一层, 分类前端 RBM 提取的特征向量, 并与正确结果比对, 进而微调整个 DBN。本文实验中的 BP 网络训练, 可参考文献[25], 利用 Sigmod 函数作为 BP 网络节点的求值函数。其训练过程如算法 3 所示。

算法 3 BP 网络的训练过程

1) 随机初始化顶层反向传播网络的参数, 设定训练步长为 N;

2) 进行前向计算, 对第 l 层的 j 单元节点, 其值为 $y_j^l(n)=\sum w_{ji}(n)y_i^{l-1}(n)$, 若神经元 j 属于输出层 ($l=L$), 则令 $y_j^L(n)=o_j(n)$, 误差 $e_j(n)=d_j(n)-o_j(n)$, d_j为正确结果;

3) 计算 δ, 将 δ反向传递用以自顶向下修正网络的权值参数, 对于输出单元:

$$\delta_j^l(n)=e_j(n)o_j(n)\ [1-o_j(n)];$$

对于隐含层单元:

$$\delta_j^l(n)=y_j^l(n)\ [1-y_j^l(n)]\sum\delta_k^{l+1}(n)w_{kj}^{l+1}(n);$$

4) 修改权值

$$w_{ji}^l(n+1)=w_{ji}^l(n)+\eta\delta_j^l y_i^{l-1}(n),$$

η为学习速率;

5) 如果 n=N, 则训练结束, 否则n=n+1, 转步骤2。

4 实验与分析

本文选取 ACE 2004 NWIRE 英文语料和 ACE 2005 NWIRE 中文语料作为实验数据, 然后按第 2.2 节给出的特征表示提取代词指代消解所需的特征语义信息, 按照 2.3 节生成训练样例及测试样例, 分别交由 Deep Learning 进行学习、测试, 形成分类器或进行分类判断, 完成指代消解这个二元分类问题。本文所用 Deep learning 学习平台采用 Bergstra 等[34]开发的 Theano 系统①。为保证结果稳定性, 我们使用 5 倍交叉验证法, 取平均值作为最终结果。

为了便于比较, 参考文献[11]所用实验数据, 表 5 给出 ACE 2004NWIRE 英文语料和 ACE 2005 NWIRE 中文语料上待消解代词的分布情况, 并且列出文献[11]所构建系统在上述语料集上得到的消解性能结果, 如表 6 所示。

表 5 待消解代词按句子距离的分布

Table 5 Distribution of pronoun anaphors over different sentence distances

指代关系跨越的句子距离	ACE 2004 NWIRE		ACE 2005 NWIRE	
	训练集	测试集	训练集	测试集
≤0	457	165	1339	684
≤1	260	73	688	320
≥2	56	33	279	169
总和	773	271	2306	1173

① 开发包下载地址: http://deeplearning.net/software/theano/

表 6　已有基准系统消解性能结果[11]

Table 6　Results of baseline system[11]

Baseline 系统	ACE 2004 NWIRE 英文语料			ACE 2005 NWIRE 中文语料		
	R/%	*P*/%	*F*/%	*R*/%	*P*/%	*F*/%
BaseLSystem_01 (基于特征向量)	65.2	78.7	71.3	64.2	78.1	70.5
BaseLSystem_02 (基于树核函数)	72.2	76.2	74.1	70.6	72.3	71.4

我们将消解对的不同词特征、语法特征、语义特征、位置特征组合后生成训练及测试样例，分别利用 DBN^i(i 表示 DBN 包含的 RBM 的层数，分别取值为 1, 2, 3)完成 3 组实验，并进行比较。第 1 组实验在不引入特征抽象值的前提下，采用不同 RBM 层数，主要验证增加 RBM 层数对消解性能的影响；第 2 组实验在使用的 RBM 层数相同的情况下，验证引入不同语义特征抽象值对消解性能的影响；第 3 组实验组合采用不同 RBM 层数及是否引入特征抽象值情况下，探索系统消解性能变化及并与 Baseline 结果作比较分析。在实验中，我们采用准确率(P)、召回率(R)和 F 系数来评价指代消解的结果。

表 7 给出第 1 组实验的结果。从表 7 可以看出：RBM 层数越多，P 值越高，即正确率越高。原因在于深度学习通过多层映射单元提取出主要的结构信息，其精确度要优于单层结构。但是，在第 1 组实验中，随着 RBM 层数的增加，系统训练时间也随之快速增加。原因是层数增加后，用于完成训练的神经网络节点增加，组合训练工作量大幅增加，这对系统平台的计算负载提出了更高要求。尽管初步实验数据表明，RBM 层数增加有利于性能提升，但是层数是否越多越好？RBM 层数与性能之间完整的关联关系如何？受限于计算平台及计算方法，目前还未进行充分验证。后续考虑拓展高性能并行计算平台，开展进一步实验加以探索。

第 2 组实验主要验证引入的不同特征抽象值对 Deep Learning 分层结构学习性能的影响，实验结果如表 8 所示。表 8 中“NL”是指不考虑特征抽象层次差异性，取值相同都为 0，其实质同第 1 组实验，即不引入特征抽象层。“YL”是指考虑特征抽象层差异性，抽象值不同，取值参考 2.2 节表 2 和 3。实验结果表明特征抽象层次的引入，对总体性能 F 值有提升，但是，影响提升的因素不同。在英文平台上，P 值有较大提升，R 值稍有降低；但是在中文平台上，P 值有降低，R 值提升，原因可能是中英文特征抽象值定义模型有差异，导致训练效果不同。拟结合实验，后续优化抽象值定义。

表 7　不同 RBM 层数平台的代词指代消解性能

Table 7　Results of pronoun resolution with different RBMs

模型	ACE 2004 NWIRE 英文语料			ACE 2005 NWIRE 中文语料		
	R/%	*P*/%	*F*/%	*R*/%	*P*/%	*F*/%
DBN^1	56.5	60.3	58.3	53.8	61.5	57.4
DBN^2	58.4	68.5	63.1	55.9	62.4	59.0
DBN^3	60.4	70.2	64.9	61.4	65.2	63.2

表 8　不同层次特征学习的代词指代消解性能

Table 8　Results of pronoun resolution with different feature levels

模型	ACE 2004 NWIRE 英文语料			ACE 2005 NWIRE 中文语料		
	R/%	*P*/%	*F*/%	*R*/%	*P*/%	*F*/%
DBN^1+NL	56.5	60.3	58.3	53.8	61.5	57.4
DBN^1+YL	56.2	66.3	60.8	58.5	59.8	59.1

第 3 组实验综合 RBM 层数及特征抽象层次两类因素，探索在指代消解中的影响，实验结果如表 9 所示。由表中数据可以看出：特征抽象层次的引入及 RBM 层数的增加，两者叠加后对性能的影响都是正面的。这也符合 Deep Learning 分层抽象学习的理论机制，即通过多层学习算法获取输入数据的主要驱动变量，进而提高学习质量。显然增加 RBM 层数可以获得多层学习；引入特征抽象层次，有利于区分输入数据的主要变量。

与已有基准系统相比较，整体性能还是有一定差距。考虑到 DBN 网络训练节点数量对训练质量的影响较大，目前所进行的实验受限于计算平台的性能，还没有实现充分测试，后续结合平台性能优化，开展充分实验，选择最佳训练参数，预计能够进一步提升系统性能。尽管如此，考虑到当前 DBN 网络基于无监督训练模式，自动化程度高，这对提升指代消解系统的自动化性能是有意义的。

表 9　基于不同 RBM 层数及不同抽象特征值的代词指代消解性能

Table 9　Results of pronoun resolution with different feature levels and RBM numbers

模型	ACE 2004 NWIRE 英文语料			ACE 2005 NWIRE 中文语料		
	R/%	*P*/%	*F*/%	*R*/%	*P*/%	*F*/%
DBN^1+NL	56.5	60.3	58.3	53.8	61.5	57.4
DBN^1+YL	56.2	66.3	60.8	58.5	59.8	59.1
DBN^2+NL	58.4	68.5	63.1	55.9	62.4	59.0
DBN^2+YL	59.7	69.2	64.1	59.2	65.1	62.0
DBN^3+NL	60.4	70.2	64.9	61.4	65.2	63.2
DBN^3+YL	65.4	70.1	67.7	64.3	66.3	65.3
BaseLSystem_01	65.2	78.7	71.3	64.2	78.1	70.5
BaseLSystem_02	72.2	76.2	74.1	70.6	72.3	71.4

5 结束语

本文结合语义信息讨论了一类代名词的指代消解问题。在使用原有语义信息的基础上，进一步提出分层泛化的语义特征表示集；同时，针对传统浅层机器学习方法的局限性，探索了基于 Deep Learning 的深层学习方法在指代消解中的应用。实验结果表明增加 RBM 训练层数可以提高系统性能；此外，引入对特征集合的抽象分层因素，也对系统性能提升产生积极作用。受限于 Deep Learning 训练平台，目前我们在 ACE 2004 NWIRE 英文语料和 ACE 2005 NWIRE 中文语料上的实验结果并没有超过现有 Baseline 系统，但也比较接近。考虑到 Deep Learning 采用的是一类无监督学习方式，后期在调整 DBN 的网络层数及网络节点数的情况下，应该还有一定的性能上升空间。此外，本文采用 Deep Learning 这一类深层机器学习替代传统浅层机器学习，实现指代消解的方式，未见文献报道。考虑到在自然语言处理中，传统浅层机器学习方法应用的广泛性，这也为进一步推动自然语言处理向利用深层语义信息，开展深度学习应用提供了有益的尝试。

下一步工作将围绕深度学习训练平台性能提升及网络训练参数优化问题开展工作。深度学习理论上具有并行特性，但是其当前采用的基于最小批处理的随机梯度优化算法，很难在多计算机中进行并行训练，因而影响训练速度，训练耗时较长。因此，结合已有研究内容，需要探索合适的平行优化算法，提升训练效率及系统性能。

参考文献

[1] McCarthy J, Lehnert W. Using decision trees for coreference resolution // Proc of the Fourteenth International Conference on Artificial Intelligence. Montreal, 1995: 1050–1055

[2] Soon W M, Ng H T, Lim C Y. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 2001, 27(4): 521–544

[3] Ng V, Cardie C. Improving machine learning approaches to coreference resolution // Proc of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, 2002: 104–111

[4] Zhou GuoDong, Su Jian. A high-performance coreference resolution system using a constraint-based multi-agent strategy // Proc of the 20th international conference on Computational Linguistics. Geneva, Switzerland, 2004: 522–528

[5] Ng V. Semantic class induction and coreference resolution // Proc of the 45th Annual Meeting of the Association for Computational Linguistics(ACL). Prague, Czech Republic, 2007: 536−543

[6] Kong F, Zhou G D, Zhu Q M. Employing the centering theory in pronoun resolution from the semantic perspective // Proc of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore, 2009: 987−996

[7] 王厚峰，何婷婷. 汉语中人称代词的消解研究. 计算机学报, 2001, 24(2): 136–143

[8] 王厚峰，梅铮. 鲁棒性的汉语人称代词消解软件学报, 2005, 16(5): 700–707

[9] 徐敏，王能忠，马彦华. 汉语中指代问题的研究及讨论.西南师范大学学报：自然科学版，1999，24(6): 633–637

[10] 王海东，胡乃全，孔芳，等. 基于树核函数的英文代词消解研究. 中文信息学报, 2009, 23(5): 33–39

[11] 孔芳，周国栋. 基于树核函数的中英文代词消解. 软件学报, 2012, 23(5): 1085−1099

[12] Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2009, 2(1): 1–127

[13] Hinton G, Osindero S, Teh Y. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554

[14] Lee T S, Mumford D. Hierarchical Bayesian inference in the visual cortex. Optical Society of America, 2003, 20(7): 1434–1448

[15] Serre T, Wolf L, Bileschi S, et al. Robust object recognition with cortex-like mechanisms. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(3): 411–426

[16] Lee T S, Mumford D, Romero R, et al. The role of the primary visual cortex in higher level vision. Vision Research, 1998, 38 (15): 2429–2454

[17] Rossi A F, Desimone R, Ungerleider L G. Contextual modulation in primary visual cortex of macaques. Journal of Neuro-science, 2001, 21(5): 1689–1709

[18] Erhan D, Bengio Y, Couville A, et al. Why does unsupervised pre-training help deep learning. Journal of Machine Learning Research, 2010, 11(3): 625–660

[19] Bengio Y, Lamblin P, Popovici D, et al. Greedy

layer-wise training of deep networks // Advances in Neural Information Processing Systems 19 (NIPS'06). Cambridge: MIT Press, 2007: 153–160

[20] Ranzato M, Poultney C, Chopra S, et al. Efficient learning of sparse representations with an energy based model // Advances in Neural Information Processing Systems 19 (NIPS'06). Cambridge: MIT Press, 2007: 1137–1144

[21] Salakhutdinov R, Hinton G E. Deep Boltzmann machines // Proc of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009. Clearwater Beach, Florida, 448–455

[22] 孙志军, 薛磊, 许阳明, 等. 深度研究综述. 计算机应用研究, 2012, 29(8): 2806–2810

[23] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classication: a deep learning approach // Proc of the 28th International Conference on Machine Learning(ICML). Bellevue, WA, 2011: 513–520

[24] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011(12): 2493–2537

[25] 陈宇, 郑德权, 赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取, 软件学报, 2012, 23(10): 2572−2585

[26] Zhou G D, Su J. Error-driven HMM-based chunk tagger with context-dependent lexicon // Proc of the 2000 Joint SIGDAT Conf on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, 2000: 71−79

[27] Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger // Proc of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, 2002: 473−480

[28] Yang X F, Su J, Zhou G D, et al. Improving pronoun resolution by incorporating coreferential information of candidates // Proc of the 42nd Annual Meeting of the Association for Computational Linguistics(ACL). Stroudsburg, 2004: 127–134

[29] Bengio Y, Lecun Y. Scaling learning algorithms towards AI // Proc of the Large-Scale Kernel Machines. Cambridge: MITPress, 2007: 321–358

[30] Hinton G E. Products of experts // Proc of the 9th Int'l Conf on Artificial Neural Networks (ICANN), Vol.1. Edinburgh, 1999: 1−6

[31] Neal R M. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report, CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993. http://www.cs. toronto.edu/~radford/review.abstract.html

[32] Hinton G E. Training products of experts by minimizing contrastive divergence. Neural Computation, 2002, 14(8):1771−1800

[33] Carreira-Perpinan M A, Hinton G E. On contrastive divergence learning // Proc of the Artificial Intelligence and Statistics (AISTATS 2005). Barbados, 2005: 33–41

[34] Bergstra J, Breuleux O, Bastien F P, et al. Theano: a CPU and GPU math expression compiler // Proc of the Python for Scientific Computing Conference (SciPy). Austin, TX, 2010: 1–7