面向生物医学实体链接的联合式学习方法

胡 宇 申德荣 聂铁铮 寇 月
（东北大学计算机科学与工程学院 沈阳 110819）

摘 要 生物医学文本蕴含着丰富的探索价值，其为生物医学工作者进行研究提供了宝贵的领域知识. 充分且高效地利用海量的生物医学文献，并从中发现重要的隐藏信息、获取专业领域知识，对生物医学研究具有重要的意义. 生物医学实体链接是对生物医学文本中的命名实体进行识别，并将表示该实体的某些字符串映射到生物医学领域知识库中对应概念. 生物医学实体链接任务通常面临两个主要的挑战：(1)自然语言描述的歧义性. (2) 自然语言文本与生物医学知识库的异构性. 传统的方法基于特征选择或规则发现，依赖于手动选择特征或定义规则，处理分阶段模型中也可能出现误差传播. 因此，本工作提出了一种深度学习和知识库相结合的实体链接方法，通过深度挖掘自然语言文本的隐藏特征，及其与知识库概念图间结构的相似性，将生物医学实体识别与实体-概念对齐两个任务进行联合式处理. 该方法旨在通过标准的生物医学知识库，自动获取生物医学实体的语义信息，挖掘生物医学实体之间的语义关系. 实验表明，该方法在实体识别与对齐方面取得了较好的效果，显著提高了任务的精确性，在实体链接核心任务上取得了超过10%的性能提升.

关键词 实体识别；实体对齐；语义分析；生物医学文本挖掘；生物医学知识库
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2022.00748

A Joint Learning Method for Biomedical Entity Linking

HU Yu SHEN De-Rong NIE Tie-Zheng KOU Yue
(*School of Computer Science and Engineering*, *Northeastern University*, *Shenyang* 110819)

Abstract Biomedical texts contain valuable domain knowledge for biomedical researchers. It is of great significance to make full use of massive biomedical literature, discover important hidden information and acquire professional knowledge from it. Biomedical entity linking is the identification of a named entity in a biomedical text and the mapping of the mention strings representing that entity to the corresponding concepts in the biomedical domain-specialized knowledge base. However, the biomedical entity linking task usually faces two major challenges: (1) Ambiguity in the description of entities by natural language. (2) Heterogeneity between natural language texts and biomedical knowledge base. Traditional methods based on feature engineering or rule discovery rely on manual feature selection or rule definition, and error propagation may also occur in the pipeline model. Therefore, this work presents an entity linking method combing deep learning and knowledge base, mining the structure similarity between natural language text and the knowledge base. The biomedical entity recognition and alignment are jointly processed aiming at automatically acquiring the semantic information of biomedical entities and mining the semantic relationship between biomedical entities through the standard

收稿日期：2020-12-29；在线发布日期：2021-05-29. 本课题得到科技部重点研发计划项目(No. 2018YFB1003404)、国家自然科学基金面上项目(No. 62172082, 62072086, 62072084)、中央高校基本科研业务费专项资金资助项目(No. N180716 010)资助. **胡 宇**，博士研究生，中国计算机学会(CCF)会员，主要研究领域为实体识别. E-mail: huyuneu@stumail.neu.edu.cn. **申德荣**(通信作者)，博士，教授，中国计算机学会(CCF)会员，主要研究领域为数据集成、数据库. E-mail: shenderong@mail.neu.edu.cn. **聂铁铮**，博士，副教授，中国计算机学会(CCF)会员，主要研究领域为实体识别、数据质量. **寇 月**，博士，副教授，中国计算机学会(CCF)会员，主要研究领域为社会网络、推荐系统.

biomedical knowledge base. Experiments show that this method achieves good results in entity recognition and alignment, and improves task accuracy significantly by achieving over 10% performance improvement on the entity link task.

Keywords biomedical entity resolution; biomedical entity alignment; semantic analysis; biomedical text mining; biomedical knowledge base

1 引　言

生物医学文本挖掘作为生物信息学研究方向之一，以简单高效的方式分析复杂的大型数据集为目标，是从海量专业领域文本中获取结构化知识的有效手段，亦为越来越多的科研工作者所关注. 生物医学实体，即代表特定现实概念的词或词组构成的术语，例如蛋白质、基因、疾病和药物名称等，是生物医学文本的重要构成部分及语义的重要承载单元. 生物医学实体链接(Biomedical Named Entity Linking，Bio-NEL)[1]，包括在特定的任务中的生物医学实体识别以及将识别出的实体与知识库中的概念对齐两个方面，作为理解生物医学文本内容的重要步骤，在生物医学文本挖掘、信息检索、语义问答、知识库构建等等领域均有着广泛而基础性的应用.

生物医学命名实体识别(Biomedical Named Entity Recognition，Bio-NER)[2]是在生物医学文本中识别生物医学命名实体的任务，如基因和基因产物、疾病名称、药物名称、化学品和物种等，是信息抽取、文本理解和知识库构建等生物医学领域应用的前置技术. 实体与知识库中概念的对齐(Biomedical Named Entity Alignment，Bio-NEA)[3]是为每一个实体分配标准化知识库的可识别标识，是生物医学实体分析领域一个非常重要的问题，其广泛应用于生物医学文本的结构化理解和语义信息检索中. 除了生物医学文本挖掘领域，在相当多的其他研究方向(如科学信息检索、数据库和本体研究)中，提供生物医学专业领域实体与知识库概念的交互映射关系也是有效信息共享的关键问题.

在通用领域(如机构名称、新闻摘要等)的实体识别过程中，命名实体通常有出现频率相对稳定、结构形式相对规范、构成规则较为统一等有利于算法识别的特点，这使得通用命名实体识别的算法性能评估可达到90%以上，具有优秀的性能[4]. 而生物医学领域中的命名实体[5]具有一系列独特的性质，它们使得该领域中的命名实体识别问题更加复杂：

(1)自然语言描述的歧义性，主要表现在：生物医学实体命名规则不统一，如Alpha UFl cells也可能被描述为UF-1 Alpha cells；实体构成模式复杂，典型的包括实体嵌套和实体边界邻接等问题，在实体粒度上也使得描述性的实体名称存在歧义性.

(2)自然语言文本与生物医学知识库的异构性. 结构化的知识库与非结构化的自然语言文本，在描述格式上有巨大的差异.

精确且自动化的文本挖掘工具可以帮助研究人员最大限度地发现并从大量文本中释放结构化信息. 传统的基于特征选择或规则发现的方法，依赖于规则设计，同时分阶段处理实体识别与对齐问题的模型可能会出现误差传播、无法有效的利用实体间隐藏信息等问题，在面对生物医学领域专有名词较多且语句构成模式较为复杂等问题时，难以设计精确度较高的算法. 因此，需要开发新的方法和工具来支持更有效和一致地提取生物医学实体及其公共数据标识符，从而促进诸如关系提取和知识库构建等一系列下游应用.

针对以上问题，本文提出了一种深度学习和知识库相结合的实体链接方法，识别生物医学实体，并将其与知识库中的描述相同现实对象的概念对齐. 该方法联合式处理实体识别与对齐，首先基于聚合神经模块的残差网络(Aggregated Neural Block-based Residual Network，ANRNet)进行实体识别任务，生成一系列可能的实体识别候选结果. 在此基础上，进一步采用基于知识库概念图路径距离的实体与概念对齐方法(Entity-Concept Alignment，EnConAli)将候选实体映射到相应概念，完成实体与知识库概念的对齐.

本工作针对实体歧义性的问题和文本与知识库异构的问题，着重进行研究，做出了以下贡献：

(1)为应对实体识别中的歧义问题，提出了一种基于ANRNet的实体识别算法. 本工作在词处理过程将单词及其近邻单词的特征合并作为当前词的特征，采用多头神经网络结构深度提取实体的边界特

征，通过学习模型识别出当前词或词组的可能候选实体.

(2)为应对文本与知识库概念图异构的问题，提出了一种基于概念图路径距离的实体-概念对齐算法(简记EnConAli)和一种基于R-P策略的改进的实体-概念对齐算法(简记EnConAli(R-P)).本工作将生物医学文本分为词、实体、概念三个层级，通过构建实体及其近邻实体的共现关系，拟合知识库中概念间的关联模式，并在概念层级实现文本词与知识库概念图的对应.

(3)本文将实体识别与对齐进行了联合式处理，以解决很多分阶段算法中误差传播的问题.采用ANRNet实体识别方法，获得一系列实体识别的候选结果；同时采用EnConAli完成实体-概念对齐.通过联合处理实体识别与实体-概念对齐任务，有效提高了生物医学实体链接的精准性.

2 相关工作

生物医学实体作为文本内容的重要组成部分和语义承载单元，一直以来都被研究人员所关注[6].生物医学实体分析相关的技术已经成为现实生物医学研究的基本工具，在生物医学中的应用正迅速从小规模的评估向大规模的应用过渡.生物医学实体挖掘[7]在专业领域文本智能挖掘、隐藏信息发现等方面均有重要应用，因而受到各领域研究者的广泛关注.在生物医学实体挖掘中，实体抽取和实体与概念的对齐是非常重要的研究方向[8]，也是以实体为中心的概念构建和文本挖掘的主要任务.

生物医学命名实体识别(BioNER)[9]是生物医学信息提取中的一项关键任务，获得了众多研究人员广泛研究.现有的大多数方法都将这一问题看作是一个序列标记任务[10]，可以通过传统的基于机器学习(ML)的模型(如隐马尔科夫模型和条件随机域[11-12])和基于特征或规则[13]来处理.但是特征设计是需要耗费大量人力的，并且具有很强的局限性，难以随着任务目标的更换而改变.基于词嵌入技术的神经网络自动提取特征[14]，在不依赖复杂特征工程的情况下，利用多层神经网络构造特征表示.基于机器学习方法，特别是基于条件随机域[15]和深度学习模型[16]的机器学习方法，取得了大量优秀的研究成果.尤其是基于双向长短时记忆网络条件随机场(Bi-LSTM-CRF)的深度学习生物医学实体识别[17]，显示出良好的结果.

然而，尽管一段时间以来，最先进机器学习算法，特别是条件随机场(CRFs)和近几年流行的深度学习模型，在实体识别领域取得了较好的效果.这一类实体识别系统的准确性仍然是面临着巨大的挑战[18].由于文献数量的快速增长，数据标注显示出极大的局限性，以及从越来越多的新加入的文本中挖掘信息的要求，该类算法面临着可扩展性问题，还有在面对未见过的文本类型或数据类型时可重用性问题.最近的研究表明，CRFs在BioNER上的高性能在推广到更多样的数据集，尤其是包括双向长短期记忆网络-条件随机场(BiLSTM-CRF)在内的基于深度学习的BioNER方法，仍然面临着可扩展性方面的挑战[19].

与实体识别相比，实体对齐任务[20]是一个更具挑战性的任务.将实体对齐到知识库概念中，采用知识库公共数据标识符[21]可以促进数据集成和重用已成为研究人员广泛共识.以前在这个子任务上的工作主要基于领域特定的字典或启发式规则[22]，并且可以实现相对较高的性能.然而，这些方法严重依赖于字典的完整性和规则的设计.因此，很难将它们应用到新数据集或将它们迁移到新域.后来，一些工作[23-24]提出了将自然语言文本和候选实体嵌入至一个公共的向量空间，进而通过评分函数(例如，余弦相似性)消除候选实体的歧义[25].近年来，基于神经网络的方法[26]在实体对齐方面同样取得了相当大的成功，其研究核心依然是算法扩展性、实体歧义、实体构成复杂等问题.随着生物医学领域高通量实验带来的大量数据的增加，这些问题变得越来越重要.利用计算机进行生物实体识别与对齐仍然是一项具有重大挑战性的任务.

可见，现有的针对生物医学实体识别与对齐的研究均存在明显不足，无法很好的满足用户的需求.为此，本文针对最常出现的实体歧义与文本-知识库异构问题进行深入的分析研究，提出了生物医学实体识别与对齐算法，进一步提高生物医学实体识别与对齐的实用性.

针对实体识别问题，本文提出了基于聚合神经模块的残差网络(ANRNet)进行实体识别任务.基于聚合神经模块的残差网络架构采用VGG/ResNets[27-28]的重复层的经验，采用分割-变换-聚合策略[29]，构建具有相同拓扑结构的神经网络模块.该模型中的任意模块均独立执行一组特征转换，将低维嵌入结果作为输出通过求和进行聚合.因此，本架构易于扩展，允许根据任务的不同扩展到任意类型的特征变换问题，

而不需要经过复杂的特殊设计.

针对实体对齐问题，本文基于假设：文本中位置相近的实体，其在知识库中对应的概念应当具有较近的路径距离. 在此基础上，采用基于概念图路径距离计算的方法将实体映射到概念图上，进行实体与概念的对齐. 并且设计了最小化图距离的词义消歧算法，通过在实体上下文语境和标准知识库的辅助下，找到多义词的正确含义，以进行实体与知识概念的映射对齐，完成标准化描述的任务.

本工作提出的方法主要针对实体歧义与文本-知识库异构等两个主要问题，将可能表示实体的单词与其上下文单词共同编码进行特征的深度建模，并在知识库中建立单词及其上下文的语义距离计算方式，以得到更精确的结果.

3　问题描述

在生物医学文本中，许多命名实体歧义源于词语的多义性，以及其修饰词或其它实体紧密相邻以至于边界难以确定等原因，对命名实体识别和对齐的任务构成较大的挑战. 传统的方法在充分捕捉实体边界的信息等方面面临较大挑战，限制了算法在应对该类任务时的性能. 本文涉及的具体符号示例如表1所示.

表1　符号含义

符号	含义
$S=(w_1,w_2,\cdots,w_K)$	单词序列S
G, C, R	生物医学知识库，概念节点及其关系
w_k	序列S的第k个单词
$w_i\|i=[t-s,t-1]\cup[t+1,t+s]$	滑动窗口2s范围内单词w_t的邻近单词
$\boldsymbol{x}_k$	第k个单词w_k的特征向量
e_k, e_{ka}	w_k对应的锚点实体及第a个候选实体
c_k, $c_{kb}\in$ C	e_k对应的锚点概念及第b个候选概念

本文将处理对象分为三个层级：词(w)、实体(e)、概念(c)，如图1所示. 词是构成句子的基本单元. 在自然语言中，实体的构成模式可以看作是单词序列组合为词组，并由若干词组构成实体. 较简单的实体是指由一个单词构成的实体；较复杂的实体是指由多个单词组成的词组序列构成的实体. 为了描述较大粒度实体的构成特征，并与本文算法所处理的数据结构相适应，本文采用多层级描述方式，并以单词及其上下文特征作为该词的基本特征.

实体是由独立的单词或词组序列表示的，描述

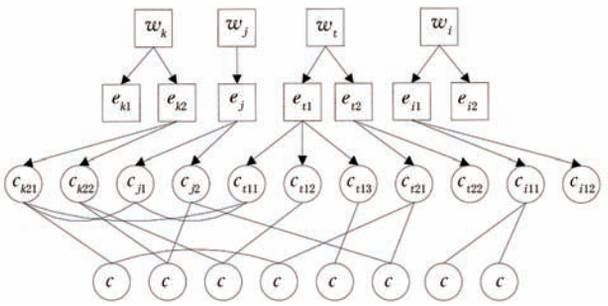


图1　实体识别与对齐实例

现实概念的词项. 即是，在由单词序列($w_1,w_2,\cdots,w_K$)表示的句子S中，元素w_k表示构成文本的单词，其中，如果若干连续单词序列($w_{k-2},w_{k-1},w_k,w_{k+1},w_{k+2}$)构成的词组表示具有明确语义的现实概念，则定义其表示实体e_k. 由于本文提出的实体识别算法用于分析句子中描述实体的单词序列的边界，因此根据描述实体的单词序列出现在文本中的位置信息，着重划分实体可能具有的类型，主要包括：

(1)简单实体：某一个单词w表示独立的实体，记为$w.e$；

(2)独立实体：实体e是由多个单词($w_{k-1},\cdots,w_k,\cdots,w_{k+1}$)组合构成，且紧邻实体$e$边界的单词$w_{k-2}$与$w_{k+2}$均不是简单实体或其它任意实体的构成部分；

(3)邻接实体：实体e由多个单词($w_{i-1},\cdots,w_i,\cdots w_{i+1}$)组合构成，且紧邻其边界位置的单词$w_{i-2}$或$w_{i+2}$是简单实体或其它具有完整语义的实体的构成部分.

概念是生物医学领域知识库(如UMLS[30])的叙词表中定义的，用于描述实体及实体间关系的标准化专业术语. 这些概念名称通常具有一致的分类体系、统一分配的标识符、标准化的单词描述形式.

文本序列表示为特征矩阵$\boldsymbol{X}\in \mathrm{Q}^{M*K}$，Q表示实数集合，该矩阵第$k$列的$M$维向量$\boldsymbol{I}_k$表示单词$w_k$的特征向量. 为了表示单词$w_k$的上下文特征，将文本序列中以每个词$w_k$为中心，滑动窗口$2s$范围内单词$w_k$的邻近上下文单词的特征向量串联，合并为$w_k$的特征向量$\boldsymbol{x}_k=(I_{k-S},\cdots,I_k,\cdots,I_{k+S})$. 以$\boldsymbol{x}_k$为列，构成新的$K$列矩阵作为文本的特征矩阵.

$$\boldsymbol{y}=\sum_{i=1}^{K} w_i \boldsymbol{x}_i$$

实体识别算法用符号J表示，本算法的主要目标是：对构成句子的每一个单词，判断其是否表示实体的一部分. 可以形式化地描述为$J\{w_1,w_2,\cdots,w_K\}$

$=\{j_1, j_2, \cdots, j_K\}$，其中$j_k=\{\mathrm{B},\mathrm{O},\mathrm{E}\}$表示对应位置的单词是否描述实体的起始边界(B)，不属于实体(O)或描述实体的结束边界(E).

实体对齐算法用符号H表示，本算法的主要目标是：将上一步识别出的实体，映射到知识库中表示概念的节点上，以实现自然语言描述的实体与标准化描述的概念进行对齐. 该过程形式化描述为$H(e_1, e_2, \cdots, e_T, G(C,R))=(c_1, c_2, \cdots c_T)$，其中$C$表示知识库$G$的概念，$R$表示知识库中概念间的关系. c_t表示实体e_t对应的概念，表示将文本描述中所包含的T个实体一一映射至概念库中的相关概念. 实体对齐算法通过实体识别和对齐任务的联合处理，纠正实体识别中可能出现的误差并完成实体链接任务.

需要指出的是，第一，通过实体识别算法对实体边界进行判断，是本文提出的实体链接算法的第一步，得到单词对应实体的候选集，在之后的实体对齐算法中，进一步通过将实体识别结果与实体对齐任务联合式处理，以得到实体链接的结果. 第二，由于实体通常是由一个单词序列表示的，且下文将实体与概念进行对齐的过程中，是在实体级进行分析处理，并不涉及某一个单独的词的处理，因此，若包含单词w_k的词组序列(w_{k-1}, w_k, w_{k+1})描述实体e_k，为了在实体对齐过程中表述方便，仍旧将该序列简记为单词w_k.

4 基于ANR-Net的实体识别

本节介绍采用基于聚合神经模块的残差网络(ANR-Net)，提取生物医学文本的隐藏信息，完成生物医学实体的识别.

4.1 生物医学实体特征构建

为了充分考虑单词上下文对实体边界的分类结果的影响，本文将单词及其一定滑动窗口内的近邻单词特征组合，为每个词构建独立的复合特征向量.

特征嵌入层接收由单词序列$(w_1, w_2, \cdots, w_K)$表示的句子S. 单词w_t的嵌入向量$\boldsymbol{x}_t$是上下文级嵌入向量[31]、语义级嵌入向量(word2vector嵌入)[32]、语法级嵌入向量(词性标记向量)[33]和字符级嵌入向量[34]的串联.

其中，字符级嵌入向量是对组成单词的字符采用*one-hot*表示，并对此矩阵进行卷积-池化操作得到的向量表示. 上下文嵌入是将当前单词及其上下文单词的字符级向量表示并联组成的矩阵作为输入，与多个滤波器进行卷积运算而得到，以滑动窗口作为交换信息的基本单元.

4.2 级联聚合神经网络模型

级联聚合神经网模型是高度模块化的网络结构，其通过级联方式堆叠一系列神经模块进行构建，任意构建块均聚合了一组具有相同拓扑结构的深度神经网络. 如图2(b)中实例所示，本文将多头神经

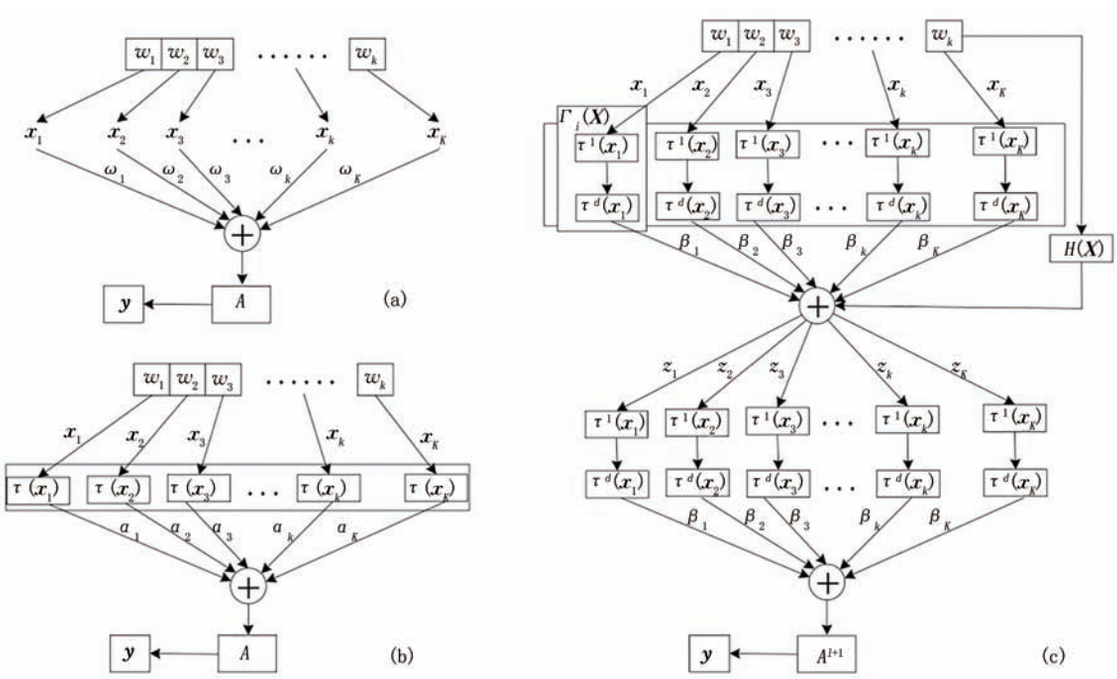


图2 ANR-Net基本架构

单元[$\tau(\boldsymbol{x}_1)\cdots\tau_i(\boldsymbol{x}_k)$]共同组成的结构称为聚合神经模块，在图2(c)所示的扩展模型中，通过级联的方式(Γ_i(X))扩展神经单元.

4.2.1　聚合神经模块

人工神经网络模型中，最简单的神经元模型可以简化为通过全连通层进行的内积操作，如图2(a)所示. 如果将内积操作定义为最简单的聚合变换：

$$y=\sum_{i=1}^{K}w_i\boldsymbol{x}_i \tag{1}$$

其中$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_K]$表示该神经元接收的$K$维输入向量，$\omega_i$表示第$i$维输入向量$\boldsymbol{x}_i$被赋予的过滤权重，该变换表示对$\boldsymbol{X}$包含的信息进行特征加权聚合，获得更高层级特征.

对于以上的简单神经元模型，可以将其理论化分解为若干基本步骤的组合：(1)分割. 将向量$\boldsymbol{X}$切片为若干个低维嵌入向量，在神经元模型中即对应一维子向量$\boldsymbol{x}_i$. (2)变换. 对全部低维表示进行加权变换，即是对低维特征的缩放操作$w_i\boldsymbol{x}_i$. (3)聚合. 通过对应位相加的方式，聚合所有变换操作后的特征向量.

通过对简单神经元模型的分析，本文尝试用通用特征变换函数$\tau_i(\boldsymbol{x}_i)$替换基础计算对象$\boldsymbol{x}_i$，如图2(b)所示. 如果该函数$\tau_i(\boldsymbol{x}_i)$描述独立的神经网络结构，则其相当于对特征向量$\boldsymbol{x}_i$进行更加复杂的按位缩放变换，也是神经网络模型在宽度上进行横向扩展. 神经聚合模型被描述为：

$$\Psi(\boldsymbol{X})=\sum_{i=1}^{K}\alpha_i*\tau_i(\boldsymbol{X}) \tag{2}$$

其中，$\tau_i(\boldsymbol{X})$为神经网络构造的特征变换函数，其以数据的第i个分量作为输入，进而通过与过滤权重α_i的内积操作，对输入特征的局部或整体嵌入的初步变换，并作为变换结果输出. K表示聚合模型中采用的变换函数的个数，在控制模型规模方面，其与简单神经元模型控制维度计数的K起到类似的作用. 本文考虑一种设计变换函数的简单方法，即所有的$\tau_i(\boldsymbol{X})$具有相同的拓扑结构，由全连通层完成特征向量变换.

4.2.2　级联聚合神经网络结构

VGG模式提出扩展重复相同结构层的策略，有助于隔离模型设计中的干扰因素，并可以轻易地扩展到任何高深度的学习模型. 本文基于此策略，在结合神经聚合模块的基础上，采用逐层堆叠级联的方式扩展结构相同的聚合模块，搭建级联聚合神经网络，如图2(c)所示.

本文将各通道内独立的转换函数$\tau_i(\boldsymbol{X})$逐层级联，通过重复堆叠这种结构，设计一种多层网络架构. 即是，神经级联聚合模型可以被描述为通过反复堆叠$\tau_i(\boldsymbol{X})$，构成的级联一系列聚合神经网络的架构：

$$\Gamma_i(\boldsymbol{X})\propto\tau_i^D(\boldsymbol{X})$$
$$\Phi(\boldsymbol{X})=\sum_{i=1}^{K}\beta_i*\Gamma_i(\boldsymbol{X}) \tag{3}$$

其中，$\Gamma_i(\boldsymbol{X})$表示变换函数，β_i表示过滤权重. $\Gamma_i(\boldsymbol{X})$以数据的第i个分量$\boldsymbol{x}_i$作为输入，通过堆叠D层基础变换函数$\tau_i(\boldsymbol{x}_i)$，并将结果与过滤权重β_i进行内积操作，对输入特征向量的局部或整体嵌入变换，并作为变换结果输出.

4.2.3　级联聚合残差神经网络结构

深度模型在网络层数增加的情况下，训练集的损失下降趋于饱和后，若网络深度持续增加，训练集损失反而会增大. 浅层网络在出现这种现象时，较于深层网络有着更好的训练效果. 残差神经网络在高低层间添加直接映射，低层特征持续向上传递，以获得相当的信息传递能力. 本文提出的聚合残差网络同样是由一系列残差块组成的：

$$y=H(\boldsymbol{X})+\sum_{i=1}^{K}\beta_i*\Gamma_i(\boldsymbol{X}) \tag{4}$$

其中，$H(\boldsymbol{X})$是残差块的直接映射部分；β_i表示过滤权重. $\Gamma_i(\boldsymbol{X})$以数据的第i个分量$\boldsymbol{x}_i$作为输入，聚合变换结果作为残差部分. 级联聚合模块有K个特征提取函数，则级联残差神经网的每一级输入都是特征直接变换$H(\boldsymbol{X})$和上一级输出$\Phi(\boldsymbol{X})$的串联拼接，模型最高层级输出为K个残差变换的平均值.

其中更深的某一层L，其与较浅层l的关系表示为：

$$\Phi(\boldsymbol{X})=\sum_{i=1}^{L-1}\sum_{i=1}^{N}\beta_i*\Gamma_i(\boldsymbol{X}) \tag{5}$$

因为对于级联聚合残差网络，L层可以表示为任意较浅l层，与高低层次间残差部分之和.

如果当前单词位于两个实体的邻接边界处，或位于实体与普通文本的边界处，以不同的输出标识作为实体边界标志.

4.3　特征连接

特征连接步骤是将各个神经模块的输出连接为特征矩阵，作为激活函数的前置部分. 本文设计了不同神经模块间输出向量的连接模式，如图1(c)所示，将这些模块中的向量叠加作为下一级神经模块的输入：

$$ConResult^{l+1}=Concate(z_1^{l+1},\cdots,z_c^{l+1},\cdots,z_C^{l+1})$$
$$ConResult_{ki}^{(l+1)}=w_{zki}^{l+1}*ConResult_{ki}^{l+1}+b_{zi}^{l+1}$$
$$A^{l+1}=activate(ConResult^{l+1}) \tag{6}$$

其中，z_c^{l+1}表示第$l+1$层的输入，$Concate(*)$表示对向量序列进行并联处理，$ConResult$表示特征连接的输出结果. 权重参数w_{zki}^{l+1}和偏置参数b_{zi}^{l+1}应用到公式所描述的连接和激活操作中，旨在将独立模块提取的多层次特征集成到同一块中. 同时，将此操作作为统一的方法来规范不同模块的输出. 该块是一个复杂的多层神经网络的子结构. A^{l+1}表示第$l+1$层的输出向量的激活函数$activate(*)$输出，再作为$SoftMax$函数的输入，得到是否表示实体的分类结果.

特征重分配步骤中，以单词w_i对应位置的输出为中心，再次将一定大小滑动窗口内的上下文作为当前位置特征向量的补充特征，共同被重新分配为w_i的特征矩阵，输入下一层级的神经模块中. 以重新分配的特征矩阵作为辅助特征，通过该特征提取结果，提高复杂边界实体识别的性能. 所有这些模块共同组成了统一的可微神经网络，该网络可以通过权值共享和反向传播算法进行联合训练.

这种设计的特征合并-抽取-重分配的上下文信息交换策略，将滑动窗口内单词特征合并，在神经网络模块内抽取特征后，再次重新分配并交换上下文信息，用于捕获嵌套实体类型的边界和跨度信息. 合并与分配机制倾向于关注并增强文本中辅助判断的部分上下文信息，并忽略或削弱不相关的信息的权重.

4.4 训练目标

级联聚合残差网络的输出向量$\boldsymbol{y}$，满足与文本正向实例的标记结果间距离尽可能的近，而负向实例的距离尽可能的远. 因此，利用基于边际的评分函数作为训练目标：

$$L=\sum_{i=1}^{K}\max\{0,\boldsymbol{y}_i-o_i+\gamma\} \tag{7}$$

其中，L表示损失函数得分，$\boldsymbol{y}_i$表示算法预测结果（章节4.2.3部分），o_i表示训练数据正确取值，γ表示调节参数. 在训练数据中使用0-1指针预测实体开始位置、结束位置，并基于开始和结束位置对构成的所有实体边界进行预测，得到实体识别的结果.

基于聚合神经模块的残差网络，采用级联聚合神经网络作为基本的计算模块. 在处理生物医学文本描述规则不统一和边界耦合复杂等问题时，可以充分利用该网络对输入特征进行局部分割，深入提取任意局部特征的隐藏信息；在局部特征聚合过程中，通过采用残差计算的方式，最大程度还原其全局特征.

在生物医学实体识别过程中，绝大部分的误差均出现在歧义部分、边界部分等自然语言的语句局部，通过该方法对局部误差进行深度分析，可以有效地提高生物医学实体识别的效果.

5 基于概念图的实体对齐方法

深层次的实体挖掘研究通常对生物医学命名实体的标注提出了更严格的需求. 生物医学文本的进一步分析需要对可能具有歧义的命名实体进行更精准更统一的标准化表示，并准确地反映实体的内部关系，从而对其语义关系进行深入的分析. 因此，本节在前文命名实体识别的基础上，基于生物医学知识库概念图，采用基于概念图路径距离的实体-概念映射算法和基于R-P策略的改进算法，将生物医学文本中名称实体对齐到知识库中的标准化概念.

5.1 知识库概念图

知识库概念图是基于统一医学语言系统（UMLS）生成的概念视图. UMLS是人工维护且规模庞大的生物医学知识库，包括术语词典、类型分类、词形词义、语义网络等诸多方面. 本文构建了更精准较小规模的实体和关系描述库（Simplified Unified Description Base，SUDB）[35]来整合和链接关键术语资源，本文称为知识库概念图. 并以此为基础，提出了将生物医学文本中名称实体对齐到标准化概念的算法.

知识库概念图SUDB由元词、元词分类和语义网络三部分组成，从实体的标准化命名方案到语义关系表达，构成了完整的知识挖掘辅助系统. 旨在提供统一的描述模式，构建生物医学名称实体之间的语义关系.

SUDB元词表：由“概念”组织，它由所有原子（术语/概念）及其变体表达式构成. 源自不同的生物医学标准，如UMLS的词汇系统. 旨在整合各种提及的概念，捕捉概念的属性，亦是简化的统一描述库的中心元素，保留了原始词汇表中描述的含义和关系，并以标准的方式进一步更新了更多的实体概念关系描述.

SUDB语义网络：每个概念属于元同义词典中的一个或多个语义类型. 这些概念通过语义网络与其他概念相联系，语义网络定义了127种语义类型和54种语义关系的编码系统.

在语义网络中，语义类型之间的边界构成了网络的基本结构，间接地表现了概念之间的语义关系. 该网络建立了“IsA”类型的层次关系和主要分为五个主要类别的非层次关系：“物理上相关”、“空间上相关”、“时间上相关”、“功能上相关”和“概念上相关”.

5.2　基于概念图路径距离的实体-概念对齐算法EnConAli

语义歧义给命名实体标准化对齐带来了挑战. 本文设计了EnConAli算法，通过在实体上下文语境和标准知识库的辅助下，找到多义生物医学实体的正确含义，以进行实体与知识概念映射对齐的任务.

EnConAli算法的基本思想是：上下文相邻的实体，在SUDB语义网络中的对应概念，应当具有较近的路径距离. 通过网络距离衡量的实体-概念关联度是将实体映射到相应概念的重要指标.

如图3所示，当前待对齐的单词w_t，其近邻单词w_k，w_i，w_j. 这种实体-概念对齐算法借鉴了MetaMap[36]的预处理方案，主要包括以下过程：

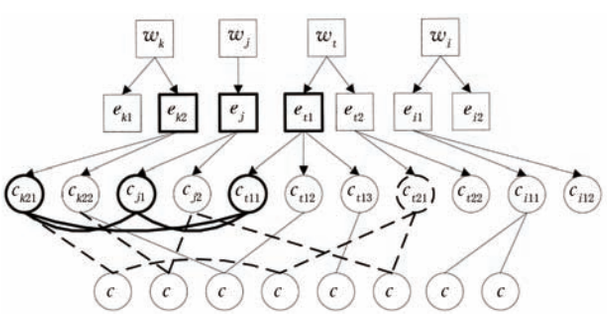


图3　基于概念图路径距离的实体对齐算法示例

1）实体标签分配：通过上文实体识别算法，并根据当前词组是否在元词表内，为生物医学文本中的词或短语单元分配标签，以指示这个标记是否引用实体. 如图3中，w_t的候选实体有(e_{t1},e_{t2})，w_k，w_i，w_j的候选实体分别为(e_{k1},e_{k2})，e_j，(e_{i1},e_{i2}).

2）候选概念生成：缩写形式等拼写变体增加了实体名称规范化的计算复杂性，这导致将自然语言文本中的名称实体与SUDB中格式标准化的实体连接起来非常困难. 因此本文对每一个标记的实体单元扩展其所有可能的变体形态（缩略词、派生变体和同义词），作为可能的候选生成. 候选概念生成是获取名称实体的扩展集的方法，将指向其任意变体的所有同义词串都检索添加，作为当前词的派生变体. 如图3中，对每一个候选实体，通过候选变量生成的方法，模糊对应到候选概念集合，如实体e_{t1}的生成的候选概念集为$(c_{t11},c_{t12},c_{t13})$，实体$e_{k2}$的生成的候选概念集为$(c_{k21},c_{k22})$.

3）候选项评估与映射构建：将所有可能的候选项进行组合，并对各组合选项综合评估，以形成尽可能准确的映射，在该步骤本文提出一种基于图距离的关联度计算来构造名称实体和概念之间的映射.

本文提出的算法基于假设：文本中位置相近的单词应当具有一定程度的关联性，并可以基于概念图的距离识别相关实体之间的关联. 如图3所示，w_t，w_j，w_k的候选实体e_{t1}，e_j，e_{k2}可能对齐的候选概念中，$c_{t11}-c_{j1}-c_{k21}$紧密相连，路径距离为3；而候选实体e_{t2}，e_{k2}，e_j的候选对齐概念，通过$c_{t21}-c-c-c_{k21}$，$c_{t21}-c-c_{j2}-c-c_{k22}$两条路径相连，因此本工作认为$w_k$，$w_i$，$w_j$可能对应概念$c_{t11}-c_{j1}-c_{k21}$的可能性更高. 即$c_{t11}$是$w_t$的对齐概念.

亦即，对于句子序列$S=(w_1,w_2,\cdots,w_K)$其标识实体的单词w，记当前单词w_t位置标识t，以t为中心的滑动窗口s；其模糊指向一系列候选实体，这些候选实体记为$e_t\in E_{wt}=[e_{t1},e_{t2},\cdots]$. 每一个候选实体可能对齐到一系列候选概念，记为$c_t\in C_{et}=[c_{t1},c_{t2},\cdots]$.

给定大小为$2s$的滑动窗口，获取相邻的单词集合$W_{i=}\{w_i\mid i=[t-s,t-1]\cup[t+1,t+s]\}$；这些单词亦模糊指向一系列候选实体，这些候选实体记为$e_i\in E_{wi}=[e_{i1},e_{i2},\cdots]$. 每一个候选实体可能对齐到一系列候选概念，记为$c_i\in C_{ei}=[c_{i1},c_{i2},\cdots]$.

算法1　基于概念图路径距离的实体对齐算法

输入：待匹配单词w_t；近邻单词w_i；概念图SUDB

输出：实体对齐结果w_k-c_k；

1. SUDB→G
2. ANR-Net(w)→[e_1,e_2…]
3. $E_{wt}=Extend(w_t)$　$e_t\in E_{wt}=[e_{t1},e_{t2},\cdots e_{ta}]$
   $Map(e_w)\rightarrow c\in SUDB$　$c_t\in C_t=[c_{t1},c_{t2},\cdots c_{ta}]$
4. $E_{wt}=Extend(w_t)$ $e_i\in E_i=[e_{i1},e_{i2},\cdots e_{ib}]$
   $Map(e_i)\rightarrow c\in SUDB$　$c_i\in C_i=[c_{i1},c_{i2},\cdots c_{ib}]$
5. FOR each a DO
6.　　FOR each b DO
7.　$ReDist(c_{ta},c_{ib})=ReDist(c_{t(a-1)},c_{i(b-1)})+Dist(c_{ta},c_{ib})$
8.　　END FOR
9. END FOR
10. let $m=minimum\{ReDist(a)\}$
11. RETURN $\{a\mid a$ in A AND $ReDist(a)=m\}$

定义1. 最短路径距离. 概念图中的任意两个点$u,v\in V(G)$的距离，定义为两者之间的最短路径的长度，标记为$G_d(u,v)$. 若存在多个点，如$(c_i,c_j,$

c_k)，其路径距离定义为

$$Dist(c_i, c_j, c_k) = G_d(c_i + c_j) + G_d(c_j + c_k) + G_d(c_i + c_k) \quad (8)$$

如果计算对象是点的集合，其描述节点到集合中路径最短的点的路径距离.

对实体e_i，通过比较其每一个候选概念与近邻概念的路径距离，综合所有可能的结果，选择出最优的实体概念匹配结果. 如算法5.1所示：

算法1的步骤1将SUDB语义网络转换为有向图G，其中每个节点都标识独立的概念以及其语义类型，且各语义类型之间的关系是节点之间的边.

步骤2和3对于指向某实体的词w，其所有的扩展词指向的实体集合为$E_w = Extend(w)$. 同时将句子实体集合E_w中的所有项通过预处理过程中的方法映射到候选的SUDB概念($Map(e_w) \rightarrow c \in \text{SUDB}$).

同时，初始化当前待匹配的单词，以及其邻居单词. 设置集合$C_{et} = \{c_i | e_t \in E_{wt},\ c_t \in Map(e_t)\}$，表示当前单词$w_t$的所有候选实体$e_t$对应的候选映射概念$c_t$；设置集合$C_{ei} = \{C_{e(t-s)}, \cdots, C_{e(t+s)} | i = [t-s, t-1] \cup [t+1, t+s] \cup c_i \in Map(e_i)\}$，表示当前单词$w_t$的所有以$2s$为大小的滑动窗口内的邻居单词$w_i$的对应实体$e_i$的候选映射概念$c_{ei}$.

步骤5-9是计算候选概念的所有可能路径距离. 在此步骤中，初始化概念间路径距离计算函数$ReDist(e_i, e_j) = \text{NULL}$，对于任意的集合$C_{et}$中的任意候选概念，及邻居实体可能对应的任意候选概念C_{et}，依次计算所有可能的两两概念组合，并计算所有概念对的最短路径距离$G_d(c_t, c_i)$. 在依次计算概念距离同时，更新路径距离函数：

$$ReDist(c_{ta}, c_{ib}) = ReDist(c_{t(a-1)}, c_{i(b-1)}) + Dist(c_{ta}, c_{ib}) \quad (9)$$

步骤10-11，对所有可能的路径组合依次进行计算，并接收语义距离最短的的概念组合为正确的映射概念. 返回当前实体概念匹配结果.

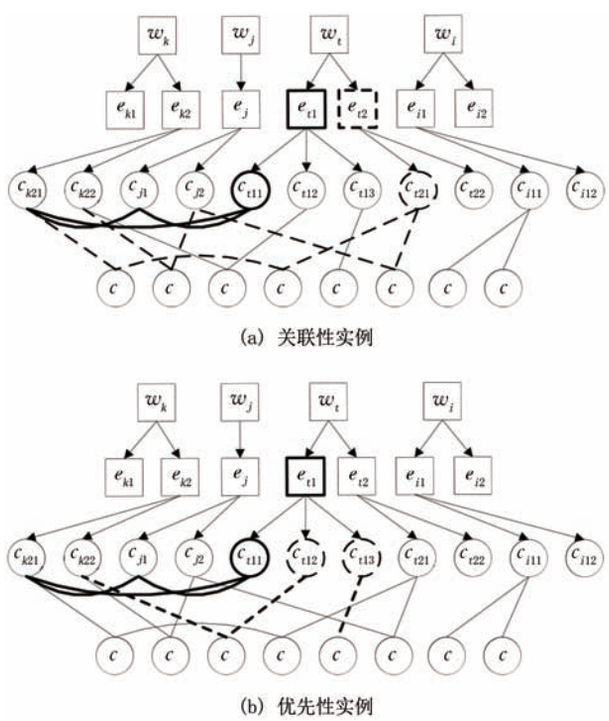
5.3 基于R-P策略的的实体-概念对齐算法EnConAli (R-P)

基于候选概念的路径距离构造实体和概念之间映射的方法，是通过迭代逐次计算的，该特性使得其依赖于初始匹配实体被正确对齐的程度. 因此，本文设计了EnConAli(R-P)算法，设定相应的度量标准，在实体识别和实体-概念对齐两个任务中，将待识别的候选实体和待对齐的候选概念进行初始选择，挑选一部分易于处理的实体与概念节点作为优先计算对象，而不止局限于随机选择的候选实体与候选概念的路径距离计算，将会有效的提高候选实体识别与候选概念项对齐两个任务联合处理的准确性，形成尽可能精准的实体链接结果.

影响实体识别和实体-概念映射对齐的两个重要因素是：

(1)候选实体与对应的实体是否有较强的关联关系. 如图4(a)所示，采用前文的实体识别与预处理方法，对词w_t的进行实体识别，得到两个可能的候选实体e_{t1}，e_{t2}. 通过定义词与实体间的关联性，并以此为依据，选择与单词w_t有较强关联的某一实体e(e_{t1}或e_{t2})作为优先处理对象.

(2)实体所对应的候选概念是否易于判别. 也就是说，实体的候选概念在对齐过程中应具有一定的辨识度，便于与其它非锚点概念进行区分. 通过计算相关概念在知识库概念图上的结构特征，赋予部分概念节点较高的优先级权重，优先计算. 如图4(b)所示，实体e_{t1}的候选概念有c_{t11}，c_{t12}，c_{t13}，通过计算每个候选概念与邻接概念的连通程度，并据此定义概念的优先级别，选择与邻接概念有较强的连通程度的概念优先计算.



(a) 关联性实例

(b) 优先性实例

图4 基于R-P策略的改进实体对齐算法示例

因此，本工作着重讨论了关联性和优先级的相关问题. 具体示例如图4所示.

在处理该问题时，本工作借鉴图的离心率相关思想. 设G是一个简单连通图，顶点集为$V(G)$，边

集为$E(G)$. 图中的任意两个点$u,v\in V(G)$的距离，定义为两者之间的最短路径的长度，记作$G_d(u,v)$. 同时定义顶点v的离心率，记为$\varepsilon(v)$，是v到图中其余各点的距离的最长路径值. 同时定义顶点v的近心率，记为$\varepsilon'(v)$，是v到图中其余各点的距离的最短路径值.

用$\deg_G(v)$表示顶点v的度. 则与图中点的离心率密切相关的另一拓扑指标：离心连通指标，其用$\xi(G)$表示，定义为

$$\xi(G)=\Sigma_{v\in V(G)}\deg_G(v)\,\varepsilon(v) \tag{10}$$

离心连通指标是基于图的连通度和离心率指标而提出的一种研究距离的指标，用以描述图内部的连通程度.

关联性. 关联性指标旨在评价候选实体与单词间的关联关系，以选择最有可能表示当前单词的实体. 由于无法直接比较候选实体与单词及其近邻单词间的关系，因此，本文通过候选实体的映射概念与近邻单词的映射概念的联通方式间接计算.

如图4(a)所示，如果某一候选实体(如e_{t1})是单词w_t对应的实体，则其某一个映射的候选概念(如c_{t11})有较大的可能是对齐的锚点概念. 该实体的当前映射概念(c_{t11})应当与邻近实体(如e_{k2},e_j)具有较强的语义联系. 亦即是说，若近邻实体模糊映射到若干概念(如e_{k2},e_j映射到概念c_{k21}, c_{j1})，其中某些候选概念项必定与当前概念(c_{t11})有较强的语义关系.

具体地，w_t可能对应实体e_{t1},e_{t2}，其近邻单词有w_i,w_j,w_k,w_t的，则其有近邻实体$e_{k1},e_{k2},e_j,e_{i1},e_{i2}$. 由于无法直接比较该候选实体$e_{t1}$与近邻实体$e_{k1},e_{k2},e_j,e_{i1},e_{i2}$间的关系，本文将近邻实体可能的模糊映射概念作为其替代值，与候选概念进行比较. 如图4所示，实体e_{t1}的与近邻实体e_{k2},e_j的候选概念通过概念图路径$c_{t11}-c_{j1}$, $c_{t11}-c_{k21}$, $c_{t12}-c-c_{k22}$直接相连，而实体e_{t2}的候选概念，通过$c_{t21}-c-c-c_{k21}$, $c_{t21}-c-c_{j2}$与近邻实体的候选概念e_{k2},e_j相连. 在两种连接模式中，前者路径距离更短且内部连通性更高，据此，认为实体e_{t1}相比于e_{t2}，与单词w_t有更近的关联性.

定义2.　实体e_i与单词w_i的关联性，结合近邻w_j及其映射实体c_j，定义为：

$$\psi(w_ie_i)=\sum_{j=i-s}^{i+s}l(h_D(e_ie_j))/\sum_{\gamma\in h_D(e_ie_j)}\zeta(c_i,c_j) \tag{11}$$

其中，对于w_i而言，$h_D(e_i,e_j)$表示对应实体e_i,e_j在概念图中路径距离小于D的所有边，这些路径称为有效路径，$l(*)$表示有效路径的个数. γ表示距离为小于固定阈值的路径，ζ表示这些路径中的最短距离.

定理2.1.　假设候选实体e_i与词w_i表示同一现实世界对象，若候选实体e_i与词w_i的关联性大于另一候选实体(除e_i外的其余候选实体，记为e_{-i})与词w_i的关联性，则实体e_i与近邻实体e_j的所有候选概念最短路径距离小于实体e_{-i}与近邻实体e_j的所有候选概念最短路径距离，即：

If $\Psi(w_ie_i)>\Psi(w_ie_{-i})$

Then $\Sigma_{j\in[i-s,i+s]}G_d(c_ic_j)<\Sigma_{j\in[i-s,i+s]}G_d(c_{-i}c_j)$.

证明.　在路径分布较均匀的情况下，可以近似地认为$l(h_D(e_i,e_j))\approx l(h_D(e_{-i},e_j))$.

若$\Psi(w_ie_i)>\Psi(w_ie_{-i})$，则$\Sigma_{j\in[i-s,i+s]}\zeta(c_ic_j)<\Sigma_{j\in[i-s,i+s]}\zeta(c_{-i}c_j)$. 同时，由于$\zeta(*)$表示路径中的最短距离，也就是说对于选中的有效路径，$G_d(c_ic_j)=\zeta(c_ic_j)$, $G_d(c_{-i}c_j)=\zeta(c_{-i}c_j)$.

即是说，$\Sigma_{j\in[i-s,i+s]}G_d(c_ic_j)<\Sigma_{j\in[i-s,i+s]}G_d(c_{-i}c_j)$.

证毕.

该定理表示，若近邻实体映射到若干候选概念，其中某些候选项必定与当前概念有较强的语义关系，那么它一定会有一部分候选概念与当前概念有较近的路径距离. 本文通过判断这种类型的路径距离辅助判断邻近实体与当前实体对应概念的关联关系.

优先性. 优先性指标用于衡量并选择候选概念中具有较高辨识度，易于对齐至相关实体的概念. 在处理待匹配概念的优先关系方面，借鉴图的离心率的思想，扩展至本问题中. 在某个实体的一系列候选概念中，其相对于由邻近实体的概念构成的概念图的离心度越大，说明其与邻近实体距离也较远，反之，其越有可能是相关实体的对齐概念.

结合图的距离计算优先关系进行权重的求解：对于实体e_t的邻近候选实体e_{k2},e_j，如果其对应的候选概念与e_t的某一候选概念(如图4(b)中的c_{t11})之间，有更短距离的概念，则其应当具有更高的优先性. 如图4(b)所示，实体e_{t1}的两个候选概念c_{t11}, c_{t12}，概念c_{t11}与近邻实体e_{k2},e_j的候选概念的概念图最短路径分别为$c_{t11}-c_{j1}$, $c_{t11}-c_{k21}$，路径长度均为1. 概念c_{t12}通过路径$c_{t12}-c-c_{k22}$，与实体e_{k2}相连，路径距离为3. 而实体e_{t2}的候选概念，通过$c_{t21}-c-c-c_{k21}$, $c_{t21}-c-c_{j2}$与近邻实体的候选概念e_{k2},e_j相连，路径长度分别为4和3. 所以认为实体概念c_{t11}相比于c_{t11}和c_{t21}，具有更高的优先级.

定义3.　w_i有候选概念c_i，w_j的候选概念集为$Map(e_j)\to C_j=[c_{j1},c_{j2},\cdots]$，结合邻近实体候选概念

图的离心率，定义候选概念的优先性：

$$w(c_i, C_j)=\max_{c_j\in C_j}\sum_{j=1}^{2s} l\left(h_{D(e_t e_j)}\right)/\varepsilon'(c_i) \quad (12)$$

通过优先性指标，选择与近邻实体的候选概念在具有广泛的联系紧密的同时，最短连接路径的距离依然较其他候选实体更短的对象，有较高的优先性，首先纳入计算序列.

定理3.1. 假设实体e_i的候选概念$c_{i1}c_{i2}$，近邻实体e_j，e_k的锚点概念分别为$c_j\in C_j$，$c_k\in C_k$，. 采用离心率约束下优先级指标作为近邻实体计算队列的排序标准，则优先级更高的候选概念与近邻实体的路径距离小于优先级更低的候选概念与近邻实体的路径距离，也就是说：

If $w(c_{i1}, C_j+C_k)> w(c_{i2}, C_j+C_k)$

Then $\Sigma_{j,k\in[i-s,i+s]}Dist\ (c_{i1}C_jC_k) < \Sigma_{j,k\in[i-s,i+s]}Dist\ (c_{i2}C_jC_k)$

证明. 若$w(c_{i1}, C_j+C_k)> w(c_{i2}, C_j+C_k)$，则$\varepsilon'(c_{i1})<\varepsilon'(c_{i2})$，即是$Dist(c_{i1}c_jc_k) < Dist(c_{i2}c_jc_k)$. 遍历滑动窗口内所有近邻实体的候选概念，则有：$\Sigma_{j,k\in[i-s,i+s]}Dist(c_{i1}c_jc_k)< \Sigma_{j,k\in[i-s,i+s]}Dist(c_{i2}c_jc_k)$

证毕.

其表示，如果在遍历邻近实体的映射概念的时候，将离心率约束下优先级较高的映射概念先纳入计算队列，可以获得更好的计算效果. 具体体现在，如果概念节点的离心率较大，则其距离当前待匹配实体的路径距离越远，其可能作为锚点概念的可能性就越小，因此离心率差距较大的邻近实体映射概念，具有相当好的区分度；与此同时，离心率约束下优先级较高的邻近实体，由于其区分度高，在后续处理中(引入第三个节点后)，会带来的误差也较小，因为距离较远的点，根据定理2($Dist(w_iC_jw_l) < Dist(w_iC_kw_l)$)可以事先排除.

依据以上思路，本文建立结合关联性与优先性关系的决策方法，提出结合关联性与优先性的综合衡量指标：

$$Score(\mathrm{R-P})=\sum_{j=1}^{n}w_j\left(\sum_{i=1}^{n}d_{ki}\psi_{ij}\right) \quad (13)$$

其表示，在选择候选匹配概念的时候，优先选择同时具有：(1)该待对齐实体与单词具有较高的关联性，(2)该实体的候选概念具有较高的识别度，相对于其它概念集合具有较高优先性.

具体过程描述如算法2所示：

步骤1-4将SUDB语义网络转换为有向图G，其中每个节点都标识独立的概念以及其语义类型，且各语义类型之间的关系是节点之间的边. 对于句中的任意词w，其所有的扩展词记为E_w. 同时将句子W中的任意单词w及其扩展词E_w映射到概念图中的候选概念($Map(e_w)=\{c|c\in G\}$). 设置集合$C_t=\{c_t \mid c_t\in Map(e_{wt}),\ e_{wt}\in E_{wt}\}$，表示当前单词$w_t$的所有扩展词$e_{wt}$对应的候选映射概念$c_t$；设置集合$C_i=\{c_i \mid c_i\in Map(e_{wi}),\ e_{wi}\in E_{wi},\ i\in[t-s,t-1]\cup[t+1,t+s]\}$，表示当前单词$w_t$的邻近单词$w_i$的所有候选实体$e_{wi}$对应的候选映射概念队列$c_i$. 初始化当前待匹配的单词$w_t$以及其邻居单词.

算法2 基于R-P策略的改进实体对齐算法

输入：待匹配单词w_t; 近邻单词w_i；概念图SUDB

输出：实体对齐结果w_k-c_k；

1. SUDB→G
2. ANR-Net(w)→$[e_1,e_2\cdots]$
3. $E_{wt}=Extend(w_t)$ $e_t\in E_{wt}=[e_{t1},e_{t2},\cdots e_{ta}]$
   $Map(e_w)$→$c\in$ SUDB $c_t\in C_t=[c_{t1},c_{t2},\cdots c_{ta}]$
4. $E_{wi}=Extend(w_i)$ $e_i\in E_i=[e_{i1},e_{i2},\cdots e_{ib}]$
   $Map(e_i)$→$c\in$ SUDB $c_i\in C_i=[c_{i1},c_{i2},\cdots c_{ib}]$
5. FOR each a DO
6. 关联性计算$R(w_te_{ta})=\psi(w_te_i)$
7. 加入关联性计算队列：Queue(R(c_{ta}))
8. 优先级计算$P(c_{ta})=\omega(c_{ta}, C_i)$
9. 加入优先级计算队列：Queue($P(c_{ta})$)
10. END FOR
11. 计算综合得分$Score(\mathrm{R-P})=\sum_{j=1}^{n}\omega_{ij}\left(\sum_{i=1}^{n}d_{ki}\psi_{ij}\right)$
12. 排序List(R-P)=Rank($Score$(R-P))
13. WHILE(List(R-P))
14. $ReDist(c_{ta},c_{ib})=ReDist(c_{i(a-1)},c_{i(b-1)})+Dist(c_{ta},c_{ib})$
15. END WHILE
16. let $m=minimum\{ReDist(a)\}$
17. RETURN $\{a \mid a$ in A AND $ReDist(a)=m\}$

步骤6-7将所有的候选实体根据相关性排序，依次加入待匹配队列. 步骤8-9选择待匹配实体队列第i个候选概念，根据所有邻近节点的候选概念进行优先性排序，并依次加入计算队列.

步骤11-12根据所有候选节点的R-P得分排序，优先计算得分高的候选实体与概念，得到置信度较高的候选概念作为后续处理的锚点概念. 步骤13-15中，对集合C_t中的任意候选概念，及第i个近邻实体w_i可能对应的任意候选概念C_i，依次计算c_t与近邻实体每一个候选概念的组合，并采纳路径距离$G_d(c_t,c_i)$最短的实体-概念对. 在依次更新i，计算概念间路径距离的同时，更新路径距离函数：$ReDist(c_{ta},c_{ib})$.

步骤16-17对所有可能的路径组合依次进行计算，并接收语义距离最短的概念组合为是正确的映射概念. 返回当前实体概念匹配结果.

该算法表示如果邻居实体在一定程度上可以用某一组概念集合表示，且当前待判断概念与邻居实体具有较强的语义关系. 则通过这种过滤标准，可以极大的提高识别精确度. 影响实体和待匹配的概念是否有正确的映射关系的两个重要因素是：该实体的邻居实体和待匹配概念的关联关系，以及多个邻居实体在匹配评价中的优先关系. 通过这两个参数筛选出易于处理的概念，作为全局实体对齐的锚点概念，推进后续处理流程. 因为对于需要通过迭代计算的方法，选择好的初始计算对象是非常重要的，而前文介绍的方法无法有效的识别那些可以被精确匹配对齐的概念.

6 实　验

实验设计包括生物医学实体识别和实体-概念对齐子任务. 生物医学实体识别任务的深度学习模型采用了反向传播算法作为训练算法. 给定一个训练集，其中的句子被标记为标准实体序列，并从选择的数据集生成训练示例用于实体识别任务. 当将每个示例发送到训练模型时，计算示例的交叉熵损失，并将梯度反向传播到各层以更新参数. 实体-概念映射任务采用所提出的算法将生物医学实体与知识库SUDB中对应的概念进行匹配，并利用生物医学实体的元义表和语义网络进行实体关系挖掘.

6.1 实验数据

GENIA Corpus[37]是以生物医学信息提取和文本挖掘为目标的标准化基准数据集. 语料库的建立是为了支持信息抽取和生物领域挖掘系统的评估. 语料库主要包含1999个Medline摘要，这些文献摘要集中于以“人体”、“血细胞”和“转录因子”为关键字的研究领域. 语料库被标注上了不同层次的语言和语义信息，主要包含分布在5大领域（蛋白质，DNA，RNA，细胞系和细胞类型）中的36种不同细分子类的实体. 其共包含400 302个单词和101 605个实体标识符.

PubAbs数据集是本文从PubMed[38]索引的生物医学研究论文摘要，作为实体识别评价的训练和测试数据. 主要针对提取选定的实体类型：基因表述和4种特定实体类型（蛋白质、DNA、RNA、细胞系和细胞类型）. 从PubMed中挑选出4146篇由大约50位医学领域的科学家撰写的文章. 然后采用半监督的方法在元叙词词典的帮助下对属于所选类型的实体进行标注. 这个新的数据集包含大约7500个经简单注释的摘要，其中包含3512个标准化基因标识符和23 652个标准化实体标识符.

6.2 实体识别评价

本文实体识别任务对比实验采用：（1）Li等人提出的Neural-Joint算法[39]. Neural-Joint算法是一种提取生物医学实体的神经关联模型. 该算法使用卷积神经网络将单词的字符信息编码到它们的字符级表示中. 然后将字符级表示、词嵌入和词性嵌入分别输入到基于双向长短时记忆的循环神经网络中学习句子中实体及其上下文的表示.

（2）Huang等人提出的LSTM-CRF算法[40]. LSTM-CRF算法以LSTM与Bi-LSTM网络作为序列标记的基本模型，并在此基础上，推导衍生出一系列基于LSTM的深度学习算法. 部分衍生算法主要包括RNN-LSTM-CRF网络和Bi-LSTM-CRF网络. 本工作采用LSTM-CRF模型作为对比方法.

（3）He等人提出的基于词嵌入的预训练算法KAWR[41]. KAWR算法将实体的先验知识从外部知识库编码到文本表示中，并引入新的知识图用于命名实体的识别. 该方法利用循环计算单元对知识库和文本数据进行预训练，对实体信息进行编码. 通过基于知识库中定义的实体关系加强上下文建模.

（4）Hu等人提出的基于融合注意力机制的实体识别算法MEID[42]. MEID算法将实体划分为包含多个词组成的复杂实体和由单个词组成的简单实体. 其使用一种融合注意机制来生成文档级特性，不仅学习同一标记实体之间的语义关联，而且更多地关注多标记实体的关系.

表2和表3描述在不同数据集上，各算法性能的对比结果. 实验结果表明，在基于PubAbs数据集的实体识别任务上，相比于Neural-Joint算法，本文提出的ANR-Net算法准确率和召回率分别提高了0.2%和2.3%，相比于LSTM-CRF算法，准确率和召回率分别提高了0.7%和0.7%. KAWR对词的向量表示进行预训练，在召回率上同样取得优秀的效果，相比于该算法，本文提出的ANR-Net算法准确率和召回率分别提高了5.6%和3.0%. MEID算法着重于处理多词项构成的复杂实体识别，在生物医学实体识别问题上，本文提出的ANR-Net算法准确率和召回率分别提高了6.8%和2.0%. 在基于GENIA数据集的实体识别任务上，相比于Neural-

Joint算法，本文提出的ANR-Net算法准确率和召回率分别提高了1.6%和2.9%，相比于LSTM-CRF算法，准确率和召回率分别提高了1.7%和2.4%. 相比于KAWR算法，本文提出的ANR-Net算法准确率和召回率分别提高了1.8%和5.2%，相比于MEID算法，准确率和召回率分别提高了2.1%和5.3%.

表2 PubAbs数据集实体识别结果评价

数据集	方法	P	R	F
PubAbs	ANR-Net	**73.4**	**80.2**	**76.6**
	Neu-Joint	73.2	77.9	75.4
	LSTM-CRF	72.7	79.5	75.9
	KAWR	67.8	77.2	71.3
	MEID	68.6	78.2	72.5

表3 GENIA数据集实体识别结果评价

数据集	方法	P	R	F
Genia	ANR-Net	**80.2**	**82.7**	**81.4**
	Neu-Joint	78.6	79.6	79.1
	LSTM-CRF	78.5	80.3	79.3
	KAWR	68.4	77.5	72.6
	MEID	68.1	77.4	72.4

总体来看，在所有的对比实验中，Neu-Joint算法同样取得了相对较优的结果，分析其原因：(1)该模型是研究者精心设计采用神经联合模型，专用于提取生物医学实体及其关系. (2)该模型采用字符级表示、单词嵌入和POS嵌入等多种前置特征提取结果，输入到基于双向长短期记忆(LSTM)的RNN中，学习实体及其在句子中的上下文表示. 其问题也是相对明显的，单元的参数为面向两个不同任务的网络所共享，因此在训练过程中会受到实体识别和关系分类任务的共同影响. 误差的传播会在一定程度上影响实验结果.

LSTM-CRF算法手动设计了各种文本与实体特征来捕获实体识别需要的隐藏信息. 不同于Neu-Joint算法是专用于生物医学实体挖掘的深度学习网络，LSTM-CRF算法包括了一系列用于序列标记的标准通用模型，其主要包括LSTM与Bi-LSTM网络，LSTM-CRF网络和Bi-LSTM-CRF网络. 这些模型可以用于POS、分块和NER任务. 相比之下，ANR-Net模型在不使用任何解析器和形态学分析工具的情况下，只使用本文提出的方法自动捕获特征，得到了较好的结果. 具体来说，我们的模型在两个任务的精度得分上都取得了较好的结果.

此外，实验结果表明，在PubAbs数据集上得到的实体识别结果通常比在GENIA数据上得到的分数一定程度上要低，这可能是由于GENIA是由领域专家良好标记的，而PubAbs数据集只是通过检索医学词汇表进行标记. 本文提出的算法在不设计特定特征来完成特定任务的情况下，获得了较好的准确率和召回分数.

GENIA数据集是精确标注等数据集，因此为了避免数据集不够精确而可能带来的误差，本工作以此数据集为测试集，进行了不同特征组合作为输入等实验. 表4显示了以不同的特征组合作为输入处理实体提取任务的实验结果. 实验结果中列出的结果显示整体性能与不同的特征输入具有一定的关系. 本实验中，手工设计的各种特征包括：(1)字符特征，即单词的字符构成，(2)语义特征，即通过word2vector进行语义嵌入，来捕获实体提取的隐藏信息. 在同一组实验中，不同算法采用相同的特性输入，各种输入特征组合设计评估结果，另外，本实验还对两种特征组合，其全部特征组合的算法在实体提取上的查全率和查全率分别为83.7%和85.3%.

表4 不同特征集对实体识别影响

特征	方法	P	R	F
组合	ANR-Net	83.7	85.3	84.5
字符	ANR-Net	72.3	**77.1**	**74.6**
	Neu-Joint	63.5	68.5	65.9
	LSTM-CRF	**72.5**	75.3	73.9
语义	ANR-Net	**75.8**	**82.8**	**79.1**
	Neu-Joint	73.6	79.3	76.0
	LSTM-CRF	72.7	80.5	76.9

显然，字符级嵌入和词级嵌入对实体识别的贡献很大. 在字符级嵌入的实验中，本文提出的算法达到更好的结果精度和召回得分，与Neural-Joint算法相比，分别为提高了精度和召回率得分8.7%和8.6%. 与LSTM-CRF方法相比，其在精度得分上要高于本文方法0.2%，而在召回率得分上本工作高于其将近1.7%. 在语义级嵌入的实验中，本文提出的算法测试结果均要好于Neural-Joint算法和LSTM-CRF算法. 与Neural-Joint算法相比，分别为提高了精度和召回率得分2.2%和3.5%. 与LSTM-CRF方法相比，在精度和召回率得分上本工作分别提高了3.1%和2.3%.

6.3 特殊类型实体的有效性

生物医学命名实体识别的实体歧义的难点主要

包括实体名称不规范和实体-实体耦合规则较为复杂，主要体现在时常出现修饰词灵活多变、嵌套实体、实体简称较多等问题. 本文针对这三种类型的实体单独测试：(1)名词修饰词(NounModifier)表示名词或其他类型的形容词作为实体修饰词出现，或者生物医学实体具有多个修饰语，而这些修饰语指的是其他实体；(2) 嵌套实体(NestedEntity)，即一个实体与另一个实体位置紧靠，即实体前后的单词属于另一个实体名称；(3) 缩写词(Abbreviation)表示实体是以缩写词的形式出现，生物医学术语是多个专有名词的标准缩写的集合.

本实验同样是以GENIA数据集为测试集，采用Li等人提出的Neural-Joint算法[39]和Huang等人提出的LSTM-CRF[40]算法作为对比算法进行比较.

表5的实验结果显示在处理特定的形成规则的名称实体时的实验结果. 整体上看，处理名词修饰词、嵌套实体和缩写词三个问题时，本文提出的方法在准确率得分上相比于另外两种算法，最好的实验结果可以分别高出2.7%、5.2%和2.0%，在召回率得分上分别高出1.4%、4.7%和4.8%. 具体地看，在名词修饰词识别的问题中，仅有Neu-Joint在准确率上高于ANR-Net约0.4%，在其余各项上，ANR-Net均取得了最优的实验结果，且全面优于LSTM-CRF方法分别为2.7%和1.4%. 缩写词识别问题同样如此，ANR-Net均取得了最优的实验结果. 实验结果显示，在面对嵌套实体的问题时，三种方法的结果均低于另外两个问题的处理结果. 也就是说，实验结果表明，与名词修饰语和缩写两种情况相比，嵌套实体并不易被识别出来. 嵌套实体识别的查准率和查全率分别低于名词修饰语实体和缩写实体的平均分6.4%和7.2%. 这可能是因为两个实体之间的嵌套关系太复杂而无法捕获. 名词修饰实体和缩写实体是相对简单的结构，因为两个名称实体之间的界限较为明显.

表5　特殊类型实体识别结果评价

数据集	方法	P	R	F
NounModifier	ANR-Net	82.2	**81.7**	**81.94**
	Neu-Joint	**82.6**	80.6	81.58
	LSTM-CRF	79.5	80.3	79.89
NestedEntity	ANR-Net	**76.4**	**78.2**	**77.28**
	Neu-Joint	71.2	75.9	73.47
	LSTM-CRF	72.7	73.5	73.09
Abbreviation	ANR-Net	**83.5**	**85.1**	**84.29**
	Neu-Joint	82.8	80.6	81.68
	LSTM-CRF	81.5	79.3	80.38

6.4　实体对齐

本文对比实验采用：(1) 经典的基于词的实体对齐算法MetaMap算法，(2) 基于图对齐的BTD算法[43]作为基准对比实验，验证本文提出的算法EnConAli的测试性能. (3)基于词嵌入的MS-LSTM算法[44]，将知识图谱用带权重的文本特征进行拓展，然后利用随机游走生成集合序列输入到skipgram模型，从而生成知识库嵌入空间. 将文本转化为知识库中的实体可以通过一个多感知监督模型(lstm+消歧机制)，将每一个文本生成一个知识库空间的点. (4) Chen等人提出的基于隐藏信息建模的文本预嵌入模型BERT-Entity-Sim[45]. 该工作在预先训练BERT的基础上，将潜在的实体类型信息嵌入到实体表示中. 同时基于BERT的实体相似度评分集成到一个最新模型的局部上下文模型中，以更好地捕捉潜在的实体类型信息.

本工作在滑动窗口大小为5的两个数据集上对算法进行测试，选择滑动窗口是为了保持可伸缩性的目标. 通过SUDB元同义词典定义的相邻词的语义类型进行拓扑距离测量，以评估相关性.

表6显示了在GENIA和PubAbs数据集上，实体-概念映射的准确率和召回率得分. 实验结果表明，与传统的方法相比，该方法能大大提高查全率和查准率.

表6　实体对齐结果评价

数据集	方法	P	R	F
PubAbs	EnConAli	**73.4**	67.2	**70.6**
	MetaMap	62.7	**69.5**	65.9
GENIA	BTD	65.9	63.2	64.3
	MS-LSTM	71.2	65.7	68.5
	BERT-Entity-Sim	72.5	67.2	69.7
	EnConAli	**80.2**	72.7	**75.4**
	MetaMap	76.5	70.3	73.3
	BTD	69.1	67.2	68.5
	MS-LSTM	77.2	70.6	73.9
	BERT-Entity-Sim	79.1	**72.9**	75.3

总体而言，该算法在两个数据集上都表现出优秀的性能. 但是，由于完整地统计全部计算结果比较困难. 因此本工作随机选择100个实体来展示统计结果. 随机选择的19个实体在SUDB元同义词典中不匹配任何概念. 这大约代表了所有实体的20%. 总计，语料库中的71个实体匹配SUDB元词库中的一个或多个概念，接近所有抽样概念的80%. 平均每个抽样的实体有3.2个映射概念.

表6的实验数据显示，在PubAbs数据集上，EnConAli算法在大多数评价指标上均优于MetaMap算法、BTD算法、MS-LSTM算法和BERT-Entity-Sim算法，相较于准确率较好的BTD算法、MS-LSTM算法和BERT-Entity-Sim算法，性能提高了大约7.5%，2.2%，0.9%，仅有在召回率上，MetaMap算法取得了69.5%的最高得分，而在准确率上，EnConAli算法最高取得了约10.7%的性能提升.

在GENIA数据集上，EnConAli算法在准确率上取得了最高的80.2%的测试结果，相较于其它算法，性能最高提升了大约11%. 在召回率测试结果上，EnConAli算法仅略低于BERT-Entity-Sim算法0.2%，分别高于其余算法2.4%，5.5%，2.1%.

为了验证实体对齐算法对于实体正确对齐到概念的性能，本工作删除了命名实体识别的预处理步骤. 如表7所示，EnConAli算法在GENIA数据集上的精度达到92.2%，高于MetaMap在GENIA测试数据集上的精度87.5%，也高于BTD算法所取得的90.1%，与MS-LSTM算法和BERT-Entity-Sim算法取得了几乎相同的测试结果. 在召回率方面，也显示出几乎同样的实验结果，分别高于MetaMap算法、BTD算法、BERT-Entity-Sim算法6.4%、4.1%、0.1%，略低于MS-LSTM算法约1%.

表7 移除实体识别过程的实体对齐结果评价

数据集	方法	P	R	F
PubAbs	EnConAli	90.4	**87.2**	**89.3**
	MetaMap	87.7	83.5	85.2
GENIA	BTD	89.3	85.1	87.2
	MS-LSTM	87.2	85.6	86.5
	BERT-Entity-Sim	**91.2**	86.5	89.1
	EnConAli	**92.2**	89.7	**91.2**
	MetaMap	87.5	83.3	85.5
	BTD	90.1	85.6	87.3
	MS-LSTM	91.2	**90.7**	91.0
	BERT-Entity-Sim	91.3	89.6	90.5

EnConAli算法在PubAbs数据集上取得了同样优于对比实验的算法性能，其准确度和召回率分别达到90.4%和87.2%，高于MetaMap在该数据集上的87.7%和83.5%，也高于BTD算法所取得的89.3%和85.1%，仅在准确率上略低于BERT-Entity-Sim算法约0.8%.

6.5 改进的实体对齐方法评价

本实验在数据集GENIA和小规模知识库SUDB上测试，并对采用经典的基于图的MetaMap算法和BTD算法作为本文提出的实体标准化对齐算法的对比算法. 在本实验中，滑动窗口大小设置为5. 特别指出的相对较小的知识库是因为生物医学知识库的构建较为困难，缺乏大规模可靠数据集；同时因为测试数据集规模相对较小，海量知识库反而可能带来更多测试困难.

本文随机选择100个实体展示统计结果. 如表8所示，实验结果表明，与传统的基于词共现的算法相比，该算法可以很大程度上提高查全率和查准率. 基于语义网络的方法的查全率和查准率均在79%以上. 本文根据关联度来限制或松弛映射条目的数量，以平衡查准率和查全率. 表中展示了EnConAli(R-P)算法的实体-概念对齐的精确度(Precision)和召回率(Recall)评分.

表8 实体对齐结果评价

数据集	方法	P	R	F
PubAbs	EnConAli(R-P)	**79.3**	**79.1**	**79.2**
	MetaMap	62.7	69.5	65.9
GENIA	BTD	65.9	63.2	64.3
	EnConAli(R-P)	**82.2**	**81.7**	**81.9**
	MetaMap	76.5	70.3	73.3
	BTD	69.1	67.2	68.5

表8的实验数据显示，EnConAli算法在所有评价指标上均优于MetaMap算法和BTD算法，相较于准确率较好的MetaMap算法，准确率性能提高了大约16.5%，召回率提高了约9.6%. 相较于准确率较好的BTD算法，准确率得分提高了大约13.4%，召回率提高了约15.9%.

同样，由于现有算法受NER误差传播的影响，实体识别过程中的误差会对结果产生影响. 为了规避误差传播的影响，验证实体对齐算法对于实体正确对齐到概念的性能，测试实体标准化结果，本文同样设置了一个没有实体识别前置任务的实验. 通过删除了命名实体识别的预处理步骤，采用准确的实体标记结果对算法进行测试. 实验结果如表9所示，EnConAli(R-P)算法在GENIA数据集上的精度达到95.2%，高于MetaMap在GENIA测试数据集上的精度达到87.5%，也高于BTD算法所取得的90.1%. 在召回率方面，也有同样的实验结果，分别高于MetaMap和BTD算法7.8%和6.1%.

EnConAli(R-P)算法在PubAbs数据集上取得了同样优于对比实验的算法性能，其准确度和召回率分别达到93.2%和91.2%，高于MetaMap在该

表9　移除实体识别过程的实体对齐结果

数据集	方法	P	R	F
PubAbs	EnConAli(R-P)	**93.2**	**91.2**	**92.5**
	MetaMap	87.7	83.5	85.2
GENIA	BTD	89.3	85.1	87.2
	EnConAli(R-P)	**95.2**	**91.7**	**93.9**
	MetaMap	87.5	83.3	85.5
	BTD	90.1	85.6	87.3

数据集上的准确度和召回率得分87.7%和83.5%，也高于BTD算法所取得的准确度和召回率得分89.3%和85.1%.相比于EnConAli(R-P)算法，改进的EnConAli(R-P)算法，同样在准确度和召回率上取得了较好的结果.

6　结论与展望

本文重点关注生物医学的边界模糊构成复杂造成的歧义性和文本与知识库异构两个主要问题，所提出的方法针对每一个单词及其上下文均通过独立的模块深度提取特征，并结合概念图挖掘其语义关系，提出了基于扩展的级联残差神经网络的深度学习模型对生物医学实体识别，并在此基础上提出了基于最短路径距离和基于关联度优先度计算的实体-概念标准化对齐方法.解决了生物文本中实体识别与标准化对齐的问题，实现较好的测试性能.

与其它算法相比，本文的方法能够有效地提高生物医学实体识别的效果.一方面，采用聚合神经网络可以深入提取任意局部特征的隐藏信息；另一方面，局部特征聚合过程中，对于出现歧义的部分，采用残差计算的方式，最大程度还原其全局特征.在实体对齐部分，基于概念图路径计算的方法可以有效的减少对标记数据的需求，在极少标注的场景下也能够完成实体对齐任务.相较于诸如社交媒体文本等领域，标记数据完善，可以将注意力集中在算法设计上.而生物医学文本分析领域，标注数据相对较少，且标注过程需要大量专业人士参与，需要花费更多的成本，因此本工作设计的算法在具有较高的实验精确度的情况下，较好的解决了该问题.

下一步工作主要考虑的方向包括扩展知识库图的关系(边)的方法，该方法将允许应用于诸如链接预测和知识库完成等任务.此外，对文本和知识库映射到统一的连续空间的嵌入方法，该方法可能具有增强文本特性并处理多义词的强大能力.将任意文本翻译成连续空间中的点的机制为更加深入的研究创造了许多机会.例如，虽然知识库的大小是有限的，但空间本身却由无数个点组成，每一个点都对应同一域的一个有效实体(但在知识库所包含的实体中并没有明确表述).如何利用这些额外的信息(例如为了用新的数据丰富知识库)可能将构成了该问题的未来的方向之一.

参　考　文　献

[1] Zheng Jin-guang, Daniel Howsmon, Zhang Bo-liang, Juergen Hahn, Deborah L. McGuinness, James A. Hendler, Heng Ji. Entity linking for biomedical literature. BMC Medical Informatics & Decision Making, 2015, 15(S-1): S4

[2] Tian Yuan-he, Wang Shen, Song Yan, Fei Xia, Min He, Kenli Li. Improving biomedical named entity recognition with syntactic information. BMC Bioinformatics, 2020, 21(1): 539-556

[3] Wang Lucy-Lu, Chandra Bhagavatula, Mark Neumann, Kyle Lo, Chris Wilhelm, Waleed Ammar. Ontology alignment in the biomedical domain using entity definitions and context// Proceedings of the Biomedical Natural Language Processing 2018 workshop. Melbourne, Australia, 2018: 47-55

[4] Wen Yan, Fan Cong, Chen Geng, Chen Xin, Chen Ming. A Survey on Named Entity Recognition//Proceedings of the 8th International Conference on Communications, Signal Processing, and Systems. Urumqi, China, 2019: 1803-1810

[5] Asif Ekbal, Sriparna Saha. Stacked ensemble coupled with feature selection for biomedical entity extraction. Knowledge Based Systems, 2013,46: 22-32

[6] John M. Giorgi, Gary D. Bader. Towards reliable named entity recognition in the biomedical domain. Bioinformatics, 2020, 36(1): 280-286

[7] Ibrahim B. Özyurt. On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining//Proceedings of the First Workshop on Scholarly Document Processing. Online, 2020: 104-112

[8] Zhou Hui-wei, Ning Shi-xian, Liu Zhe, Lang Cheng-kun, Liu Zhuang, Lei Bi-zun. Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes. BMC Bioinformatics, 2020 ,21(1): 35-50

[9] Denis Newman-Griffis, Ayah Zirikly. Embedding Transfer for Low-Resource Medical Named Entity Recognition: A Case Study on Patient Mobility//Proceedings of the Biomedical Natural Language Processing 2018 workshop. Melbourne, Australia, 2018: 1-11

[10] Hong S. K., Lee Jae-Gil. DTranNER: biomedical named entity recognition with deep learning-based label-label transition model. BMC Bioinformatics, 2020, 21(1): 53

[11] Tang Zhuo, Jiang Lin-gang, Li Yang, Li Ken-li , Li Ke-qin. CRFs based parallel biomedical named entity recognition algorithm employing MapReduce framework. Cluster Computing, 2015, 18(2): 493-505

[12] Robert Leaman, Lu Zhi-yong. TaggerOne: joint named entity

recognition and normalization with semi-Markov Models. Bioinformatics, 2016, 32(18): 2839-2846

[13] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, Juliane Fluck. ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics. 2005,6(S-1):1-9

[14] Minsoo Cho, Jihwan Ha, Chihyun Park, Sanghyun Park. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. Journal of Biomedical Informatics, 2020, 103: 103381-103389

[15] Li Yang, Zhou Yan-hong. Exploring feature sets for two-phase biomedical named entity recognition using semi-CRFs. Knowledge Information System. 2014, 40(2): 439-453

[16] Hyejin Cho, Hyunju Lee. Biomedical named entity recognition using deep neural networks with contextual information. BMC Bioinformatics, 2019, 20(1): 735-746

[17] Nathan Greenberg, Trapit Bansal, Patrick Verga, Andrew McCallum. Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 2824-2829

[18] Minsoo Cho, Jihwan Ha, Chihyun Park, Sanghyun Park. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. Journal of Biomedical Informatics, 2020, 103: 103381

[19] Li Jing, Sun Ai-xin, Han Jiang-lei, Li Chenliang. A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering, 2020, 99:1-20

[20] KaewphanSuwisa, HakalaKai, MiekkaNiko, SalakoskiTapio, GinterFilip. Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modeling. The Journal of Biological Databases and Curation, 2018(2018): bay096-106

[21] Wu Y, Liu X, Feng Y, Wang Z, Yan R, Zhao D. Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao, China 2019: 5278-5284

[22] Richard T-H Tsai, Cheng-Lung Sung, Hong-Jie Dai, Hsieh-Chuan Hung, Ting-Yi Sung, Wen-Lian Hsu. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. BMC Bioinformatics. 2006 7(S-5) :11-25

[23] Zhu M, Busra C, Parminder B, Chandan K. R. LATTE: Latent Type Modeling for Biomedical Entity Linking//Proceedings of The Thirty-Second Innovative Applications of Artificial Intelligence Conference. New York, USA, 2020: 9757-9764

[24] Li X, Feng J, Meng Y, Han Q, Wu F, Li J. A Unified MRC Framework for Named Entity Recognition//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 5849-5859

[25] Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, Dong Huang. CNN-based ranking for biomedical entity normalization. BMC Bioinformatics, 18(S-11): 79-86

[26] Italo Lopes Oliveira, FiletoRenato, SpeckRené, Luís Paulo F. Garcia, Diego Moussallem, Jens Lehmann. Towards holistic Entity Linking: Survey and directions. Information System, 2021, 95: 101624-101640

[27] Karen S, Andrew Z. Very Deep Convolutional Networks for Large-Scale Image Recognition//Proceedings of 3rd International Conference on Learning Representations. San Diego, USA, 2015:1-14

[28] He Kai-ming, Zhang Xiang-yu, Ren Shao-qing, Sun Jian. Deep Residual Learning for Image Recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA,2016: 770-778

[29] Xie Sai-ning , GirshickRoss B., DollárPiotr, Tu Zhuo-wen, He Kai-ming. Aggregated Residual Transformations for Deep Neural Networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5987-5995

[30] Liz Amos, David Anderson, Stacy Brody, Anna Ripple, Betsy L. Humphreys. UMLS users and uses: a current overview. Journal of the American Medical Informatics Association,2020, 27(10): 1606-1611

[31] Zainab Awan, Tim Kahlke, Peter J. Ralph, Paul J. Kennedy. Chemical Named Entity Recognition with Deep Contextualized Neural Embeddings//Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. Vienna, Austria,2019: 135-144

[32] Tomás Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space// Proceedings of the 1st International Conference on Learning Representations. Scottsdale, USA, 2013: 1-9

[33] Bai Yu-xuan, Wang Yu, Xia Bin, Li Yun, Zhu Zi-ye. Adversarial Named Entity Recognition with POS label embedding//Proceedings of the 2020 International Joint Conference on Neural Networks. Glasgow, UK,2020: 1-8

[34] Mourad Gridach. Character-level neural network for biomedical named entity recognition. Journal of Biomedical Informatics, 2017, 70: 85-91

[35] Brian Walsh, Sameh K. Mohamed, Vít Novácek. BioKG: A Knowledge Graph for Relational Learning On Biological Data// Proceedings of the 29th ACM International Conference on Information and Knowledge Management. Dublin, Ireland, 2020: 3173-3180

[36] Christopher Agrafiotes, Avi Arampatzis. Augmenting Medical Queries with UMLS Concepts via MetaMap//Proceedings of the 25th Text REtrieval Conference. Gaithersburg, USA, 2016 (500-321):1-6

[37] Nigel Collier, Hyun Seok Park, Norihiro Ogata, Yuka Tateishi, Chikashi Nobata, Tomoko Ohta, Tateshi Sekimizu, Hisao Imai, Katsutoshi Ibushi, Jun'ichi Tsujii. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers//Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics. Bergen, Norway, 1999: 271-272

[38] Lu Zhi-yong. PubMed and beyond: a survey of web tools for

searching biomedical literature. The Journal of Biological Databases and Curation，2011(2011)：baq036

[39] Li Fei，Zhang Mei-shan，Fu Guo-hong，Ji Dong-hong. A neural joint model for entity and relation extraction from biomedical text. BMC Bioinformatics，2017，18(1)：1－11

[40] Huang Zhi-heng，Xu Wei，Yu Kai. Bidirectional LSTM-CRF Models for Sequence Tagging. Computer Science，2015：1-10

[41] He Qi-zhen，Wu Liang，Yin Yida，Cai He-ming. Knowledge-Graph Augmented Word Representations for Named Entity Recognition//Proceedings of the Innovative Applications of Artificial Intelligence Conference. New York，USA，2020，34(05)：7919-7926.

[42] Anwen Hu，Zhicheng Dou，Jian-Yun Nie，Ji-Rong Wen. Leveraging Multi-Token Entities in Document-Level Named Entity Recognition//Proceedings of the AAAI Conference on Artificial Intelligence. New York，USA，2020，34(05)，7961-7968.

[43] Wessam Gad El Rab，Osmar R. Zaïane，Mohammad El-Hajj. Biomedical text disambiguation using UMLS//Proceedings of the Advances in Social Networks Analysis and Mining. Niagara，Canada，2013：943-947

[44] Dimitri Kartsaklis，Mohammad Taher Pilehvar，Nigel Collier. Mapping Text to Knowledge Graph Entities using Multi-Sense LSTMs//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels，Belgium，2018：1959-1970

[45] Shuang Chen，Jinpeng Wang，Feng Jiang，Chin-Yew Lin. Improving Entity Linking by Modeling Latent Entity Type Information//Proceedings of the AAAI Conference on Artificial Intelligence. New York，USA，2020，34(05)，7529-7537

Hu Yu，born in 1990，Ph. D. candidate. His research interests include entity identification.

Shen De-Rong，Ph. D. Her research interests include data integration，database.

Nie Tie-Zheng，Ph. D. His research interests include entity identification and data quality.

Kou Yue，Ph. D. Her research interests include link predication，social media and recommendation system.

Background

Biomedical entity linking is gradually changing from feature-based method to deep learning-based method. At the same time，the research interest of experts in the field of bioinformatics is shifting to designing better neural network topologies for better learning about the representation of biomedical entities. With the increase of parameters，it becomes more and more challenging to design a reasonable and extensible architecture.

The linking of entity and concept in knowledge base and the allocation of standardized identification in knowledge base for each entity is a very important issue in biomedical entity analysis，which is widely used in structured understanding of biomedical texts and semantic information retrieval. Biomedical entities alignment problem，however，by the entity to the knowledge base of the concept of mapping is still faced with enormous challenges：the first is that the main difficulties of the ambiguity of natural language description，and entity-concept heterogeneous in alignment problem to be solved urgently. Some deep learning scholars have been working on architectural patterns with some generality in the expectation that different tasks can be handled more easily. A series of models based on ResNets designs have shown that well-designed topologies can achieve convincing accuracy with low theoretical complexity.

In this work，entity linking is regarded as a word sense disambiguation problem，and multiple concepts that map ambiguous entities to knowledge base are disambiguated to find the correct solution that represents the current entity meaning. Focusing on the issue of entity alignment in the biomedical domain，this work argues that combining unsupervised learning with an established knowledge base should be most effective.

In this study，it is assumed that adjacent entities in the text should have some degree of correlation，and for the concept of proper alignment of neighboring entities，such correlation can be measured by their path distance on the knowledge base concept map. Based on this hypothesis，this paper proposes a word sense disambiguation algorithm based on concept map path distance and an improved algorithm based on R-P strategy to complete the concept alignment of biomedical entity- knowledge base.

This research was partially supported by grants from the National Key R&D Program of China (No. 2018YFB1003404)，the National Science Foundation of China (Grant No. 62172082，62072086，62072084)，Chinese Universities Scientific Fund (Grant No. N180716010).