文章编号：1003-0077（2017）00-0000-00

融入丰富信息的高性能神经实体链接

李明扬 姜嘉伟 孔芳*
（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘要：歧义的存在使得实体链接任务需要大量信息的支撑。已有研究主要使用两类信息，即实体表述所在文本的信息和外部的知识库信息。但已有研究对信息的使用存在以下两个问题：首先，最新通用知识库规模更大、覆盖面更广，但目前的实体链接模型却未从中受益，其性能没有相应的提升；其次，表述所在的文本信息既包含表述所处的局部上下文信息，也包含文本主题之类的全局信息，文本自身信息的利用率还需进一步提升。针对第一个问题，本文给出了一个融合文本相关度和先验知识的实体候选集抽取策略，提高了对知识库中有效知识的提取；对第二个问题，文本给出了一个融合局部和全局信息的自注意力机制与高速网络相结合的神经网络实体链接框架。在 6 个实体链接公开数据集上的对比实验表明了本文提出方案的有效性，在最新的通用知识库上本文给出的实体链接模型取得了目前最好的性能。

关键词：实体链接；自注意力机制；高速网络

中图分类号：TP391 **文献标识码：**A

Towards Better Neural Entity Linking via Rich Information

LI Ming Yang , JIANG Jia Wei, KONG Fang
(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: The existence of ambiguity makes the entity linking task needs a large amount of information. Previous research mainly uses two types of information, i.e., the information of the text containing the given mention and the external knowledge base. There are still two issues should be addressed. Firstly, current entity linking models have not benefited from the latest knowledge base, which has larger scale and wider coverage. Secondly, the text contains rich information including local context information of the mention and global information such as text topic. The combination approach of local and global information can be further improved. For the first problem, an entity candidate extraction approach considering both text relevance and prior knowledge is proposed to get the effective entity candidate set. For the second problem, a neural network with self-attention and highway network is proposed to represent both local and global information for entity linking. Experiments on six public datasets of entity linking show the effectiveness of our proposed approach. Furthermore, our system achieves the state-of-the-art performance using the latest general knowledge base.

Key words: Entity Linking; Self-attention Mechanism; Highway Network

0 引言

实体链接（Entity Linking，EL）是指将文本中的表述（mention）链接到知识库（Knowledge base，KB）中相应实体（entity）的任务。作为自然语言理解任务的关键步骤，实体链接在问题回答（Question answering[1]）、语义检索（Semantic search[2][3]）和信息抽取（Information extraction[4]）等任务中起着基础性的作用。实体链接任务具有挑战性，因为文本中的表述本身具有较强的模糊性，单从表面形式上很难与知识库中的实体相链

基金项目：国家自然科学基金面上项目（61876118）；国家自然科学基金人工智能应急管理项目(61751206）
*通信作者：kongfang@suda.edu.cn

接，再加上表述存在一词多义的情况（例如在上下文中的“中国”一词，在不同语境中既可以表示“中国（国家）”，也可以表示“中国乒乓球队”，还可以表示“中国足球队”等），这进一步增加了实体链接任务的难度。

实体链接任务通常分为两个主要阶段：候选集生成和候选实体消歧。候选集生成是为文本中的每一个表述获取一组知识库内相关的实体集合作为候选集。候选实体消歧是在已有候选集上，通过对候选实体进行相关性排序来获取最佳的实体链。本文从三个方面对实体链接模型进行了优化：首先，给出了一个融合相关度和先验知识的候选集生成方法，提高了候选实体的覆盖率；其次，给出了一个融合局部上下文信息和全局篇章级主题一致性信息的候选实体消歧策略；最后，在进行局部上下文信息表征时，提出了自注意力机制与高速网络相结合的表征策略。

在 6 个公开数据集上的实验证明，本文给出的候选集选取策略能更好地从知识库中获取高效的候选集，融合全局和局部信息的实体消歧模型能更好地完成链接实体的选择，在 Wiki_2018 知识库上取得了目前最好的实体链接性能。

1 相关研究

早期的实体链接研究仅关注于如何将文本中抽取到的实体连接到知识库中，忽视了位于同一文档的实体间存在的语义联系。目前，为了避免手工特征的设计和减少对语言学知识的依赖，实体链接任务逐渐转向完全使用深度学习，借助神经网络强大的特征抽象和泛化能力，来直接学习文本中潜在的相关信息的基本特征及其组合。

候选集生成方面，已有研究均通过统计维基百科以及其他公开知识库中表述和实体的共现情况来解决。但这种方法存在明显的由于不分领域、不设上限的统计共现情况而导致的潜在的候选集中包含大量噪音的问题。近年来实体链接任务所使用的知识库几乎都是 2014 年的维基百科（Wiki_2014），随着维基百科 2018 版本的发布，我们将目前主流的两个实体链接模型切换到了更大更全的 Wiki_2018 上，性能并未如预期那样的有所提升，这也说明候选集选取确实存在噪音过多的问题。

候选实体消歧方面，已有研究已经由使用简单的启发式规则过渡到将单词和实体用连续空间中的低维向量表示，表述和实体的特征自动从数据中学习，最后通过近似算法来对候选实体综合排名。代表性的研究有：Sun[5]在 2015 年将表述和实体以及上下文进行嵌入式表示，并通过卷积神经网络提取特征并计算表述和实体的相似度进行链接。2016 年，Francis-Landau[6]在 Sun[5]的基础上堆叠去除噪声的自动编码器分别学习文本的上下文和实体的规范描述页面，提升了链接性能。Ganea and Hofmann[7]在 2017 年提出了用局部和全局模型结合的方式进行链接，并且在局部模型中提出了使用软注意力（Soft attention）和硬注意力（Hard attention）来筛选上下文中的单词，从而进一步提升了链接性能。2018 年，为了解决相同子句中相同的两个表述链接到知识库中不同的实体的情况，Le and Titov[8]在 Ganea and Hofmann[7]的基础上对表述进行关系建模并以特征的形式加入全局模型中，取得了目前最好的性能。

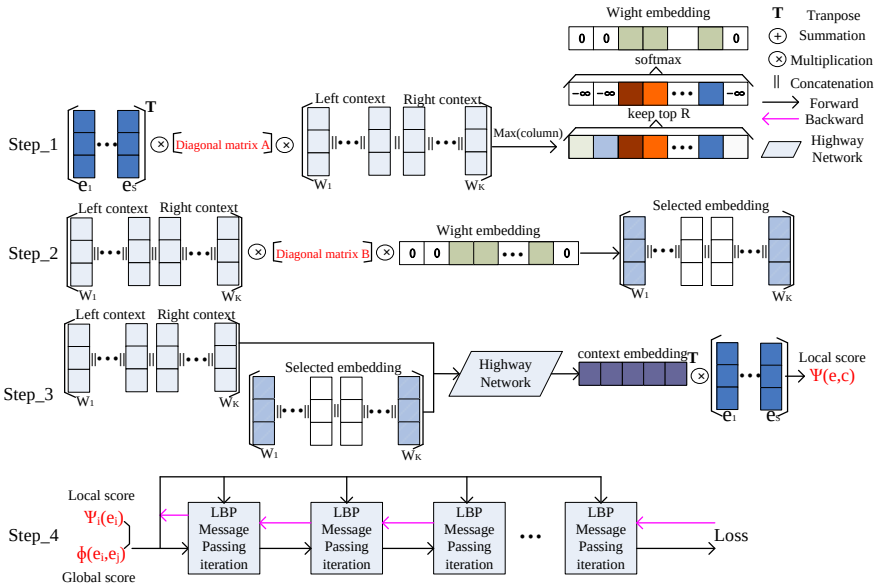

图 1 NSEL 模型

2 实体链接框架

在 Ganea and Hofmann[7]和 Le and Titov[8]的工作基础上，本文提出了神经自注意力实体链接（Neural Self-attention Entity Linking，NSEL）模型，图 1 给出了 NSEL 模型的完整框架，从中可以看到，NSEL 模型包含局部模型和全局模型两部分，步骤 1~3 构成了局部模型，第 4 步是全局模型。在介绍模型的各个构成部分之前，我们首先通过一个示例说明一下 NSEL 模型的工作流程。

首先，输入序列“France beat Croatia 4-2 in the 2018 World Cup final.”，其中下划线标出的即为待链接的表述。以“France”为例，在 step_1 中通过计算 France 的上下文与其候选实体的相关度对上下文进行筛选，相关的单词会被赋予一定的权重，而无关的单词将不予考虑。step_2 将使用权重矩阵和上下文的嵌入式表示计算出 France 的有意义单词表征。step_3 使用高速网络对 step_2 生成的单词表征和通过多头自注意力机制捕获的序列表征进行桥接，进一步丰富上下文表征，并计算出最终的局部得分。step_4 将通过对 France、Croatia 和 2018 Word Cup 等多个实体进行相互制约，在这 3 个表述具有主题一致性的假设下计算出全局得分。再综合考虑局部得分和全局得分来确定表述对应的实体。

下面我们逐个介绍 NSEL 模型中的各个组成部分。

2.1 单词和实体的嵌入式表示

在编码阶段，原始数据通过查找字向量表转化为字向量序列。对于文本中的上下文单词，我们使用预先训练完成的 GloVe[9]词向量，该词向量的维度为 300 维。对于候选集中的实体，与文本中上下文单词不同的是：文本中上下文单词可以具有多种语义，具有泛化性，而候选集中实体的语义必须只能含有一种语义，具有特定性。同时，为了更好的表现出排名靠前的实体（尤其是排名第一的实体）与其他候选实体之间的差别，我们效仿 Ganea and Hofmann[7]等人的工作，对所有语料中的候选实体重新训练它们的嵌入式表示。

训练实体的嵌入式表示时，我们先用均值为 0、标准差为 1 的正态分布随机初始化实体向量，将选择实体所在的维基百科页面文章内“Description”部分的单词作为正采样，而将随机从其他页面选出的单词作为负采样。在每次迭代中，我们会进行 20 次正采样和 5 次随机负采样，期望的结果是：与随机负采样的单词向量相比，正例单词的向量在点积相似度上更接近于实体的向量。通过如下的策略进行实体嵌入式表示的优化，最终生成语料中所有候选实体的表示。

我们采用 Ceccarelli[10]等人发布的实体关系数据集来验证和测试我们重新训练得到的实体嵌入式表示。该数据集的验证集和测试集分别包含 3673 和 3319 个查询对，每个查询对由一个目标实体和 100 个候选实体组成，以及 0/1 标签表示两个实体是否有关系。关系越紧密的候选实体将会排在其他实体的前面，所以本文使用标准化贴现累积收益（Normalized Discounted Cumulative Gain[11]，NDCG）和平均精度均值（Mean Average Precision[12]，MAP）来衡量候选实体的相关性排序质量（见表 3）。MAP 算法只能判断出实体之间是否有关系，无法描述实体间的关联程度，而 NDCG 算法将实体间的关系改进为分等级的相关度，评估时我们设置的等级为：1、5、10。NDCG 算法和 MAP 算法的计算公式如式（1）~（2）。

$$N(n) = Z_n \sum_{j=1}^{n} \frac{\left(2^{r(j)} - 1\right)}{\log(1+j)} \quad \#(1)$$

$$MAP = \bar{p}(r) = \sum_{i=1}^{Nq} \frac{Pi(r)}{Nq} \quad \#(2)$$

2.2 多知识库候选集生成及优选

考虑到生成的候选集的完备性和覆盖率，我们通过统计维基百科（Wikipedia）、大型网络语料库 Crosswikis（Spitkovsky and Chang[13]）和 Yago 知识库（Hoffart et al.[14]）中$< mention - entity >$共现的次数来生成每个表述的候选集以及表述与每个候选实体之间的先验概率$\hat{p}(e_i|m)$。若在上述 3 种不同的资源中得到了相同$< mention - entity >$的不同共现次数，则取共现次数最大的值。先验概率$\hat{p}(e_i|m)$的计算公式如式（3）所示。

$$\hat{p}(e_i|m) = \frac{\#links\ with\ m\ that\ point\ to\ e_i}{\#all\ links\ with\ m} \quad \#(3)$$

其中，m表示 mention，e_i表示 mention 的第i个候选实体。

在实验过程中，考虑到运行内存和算法时间复杂度的限制，同 Ganea and Hofmann[7]和 Le and Titov[8]的工作，我们先选出按先验概率$\hat{p}(e_i|m)$排前 30 位的候选实体，再从中选出 7 个候选实体。不同的是，我们通过以下 2 个步骤筛选出 7 个候选实体作为模型的输入：

1）按上下文相关度筛选出前 3 个实体；

2）按先验知识筛选出 4 个实体。

步骤 1 会先从前 30 位候选实体的基础上选出与表述上下文最相关的 3 位候选实体，具体过程如图 2 所示。

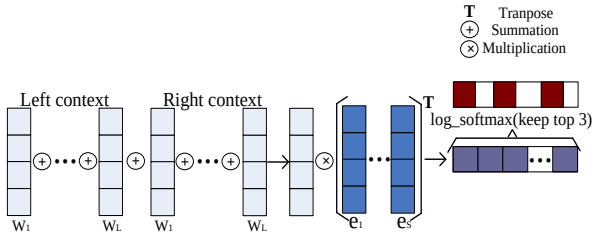

图 2 候选实体生成

其中，表述的左右单词数设置L为 25，在上下文单词求和之前会简单的去除一些噪音词，如：1）停用词表中的单词；2）数字类的单词；3）单词的长度为 1。

步骤 2 会在步骤 1 得到的 3 个候选实体的基础上按先验概率的排名继续补充 4 个候选实体，即得到了兼顾上下文相关度和先验概率的无重复实体候选集。

在接下来的两个部分中，我们将介绍本模型的关键组件：局部模型和全局模型。

2.3 局部模型

局部模型是通过表述周围的上下文信息为每个表述生成每个候选实体与上下文的相关性得分。考虑到上下文中并不是所有单词都能提供有用的信息，会存在无信息的单词(停顿词、定冠词等)因为出现频数、出现位置等因素获得较大的得分，往往会对$< mention - entity >$产生负面影响。所以我们提出假设：若上下文中的单词是与表述相关，则它至少与该表述的一个候选实体是强相关的。

鉴于以上的假设，我们通过计算表述周围单词与该表述的所有候选实体之间相关性得分，并选出每个单词与所有候选实体最高得分来筛选出相关的上下文单词。如图 1 中的 step_1 所示，选出得分排名前R个单词后，将其他单词的得分值设为$-\infty$，以便在$softmax$操作后得到的权重为 0，即忽视该单词。数学定义如下：我们将每个表述记做m，表述的候选实体集为$\Gamma(m)$，候选实体$e \in \Gamma(m)$。表述的上下文单词记做$c = \{w_1, w_2, \dots, w_K\}$，简记每个$w \in c$，$x_w$表示上下文单词的嵌入式表示。则 step_1 的计算公式如式（4）～（6）所示。

$$\mu(w) = \max_{e \in \Gamma(m)} x_e^T A x_w \ \#(4)$$

$$\bar{c} = \{w \in c | \mu(w) \in topR(u)\} \#(5)$$

$$\beta(w) = \begin{cases} \dfrac{\exp[\mu(w)]}{\sum_{v \in \bar{c}} \exp[\mu(w)]} & if\ w \in \bar{c} \\ 0 & otherwise \end{cases} \#(6)$$

step_2 是在 step_1 得出上下文单词权重的基础上得出筛选后的上下文嵌入式表示$Select_x_w$，如式（7）所示。

$$Select\_x_w = \sum_{w \in \bar{c}} \beta(w) B x_w \ \#(7)$$

由于文本内潜在的序列化信息也相当重要且未被使用，step_3 首先使用自注意力机制捕获上下文潜在的序列信息，再通过高速网络对筛选单词信息和序列信息进行桥接（见 2.4 节）。最后将生成的上下文表征$Cont$与候选实体进行相似度计算得到局部模型的$< mention - entity >$得分$\Psi(e, c)$，如式（8）～（9）所示。

$$Cont = Highway\_Network(x_w, Select\_x_w) \#(8)$$

$$\Psi(e, c) = \sum_{w \in \bar{c}} Cont * x_e^T \ \#(9)$$

2.4 融入自注意力机制的高速网络的应用

多头注意力（Multi-Head Attention[15]）机制可以从多角度、多层级的视角提取更多的文本自身的特征，本质是将输入向量Q, K, V通过参数矩阵映射后再做放缩点积注意力（Scaled Dot-Product Attention）机制，并将这个过程重复做h次，最后将结果进行拼接，从而获得较全面的特征信息。

首先介绍放缩点积注意力机制，其本质上是使用点积进行相似度计算。计算方式如式（10）所示。

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \#(10)$$

其中，Q, K, V均为向量形式，且$Q \in \mathbb{R}^{n \times d_K}, K \in \mathbb{R}^{m \times d_K}, V \in \mathbb{R}^{m \times d_V}$，$d_K$表示$Q, K$的第二维度。当$Q, K, V$为同一个输入序列时，该 attention 即为自注意力（Self-attention），挖掘序列内部的信息。

在此基础上，考虑到一个 attention 机制无法从多角度、多层面的捕获到重要的特征，所以需要使用多头注意力（Multi-Head Attention）机制。图 4 给出了多头注意力机制的计算方式，该 attention 的计算如式（11）～（12）所示。

$$head_i = Attention(QW_i^Q, KW_i^k, VW_i^v) \#(11)$$

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h) \#(12)$$

其中，$W_i^Q \in \mathbb{R}^{d_K \times \tilde{d}_K}, W_i^k \in \mathbb{R}^{d_K \times \tilde{d}_K}, W_i^v \in \mathbb{R}^{d_v \times \tilde{d}_v}$，$Concat$表示将每次的结果进行拼接。

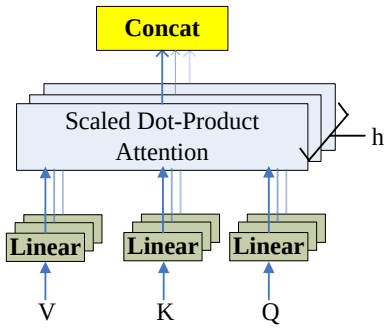

图 3 多头注意力机制

高速网络（Highway Networks[16]）是一种能够在信息传递之间进行平滑切换的神经网络。本文使用高速网络对 step_2 筛选出的有意义单词表征和多头自注意力机制捕获的文本序列信息进行桥接，让模型自己训练出两者的结合比例，最后得到更全面、丰富的上下文的嵌入式表示$Cont$，如图 4 所示。

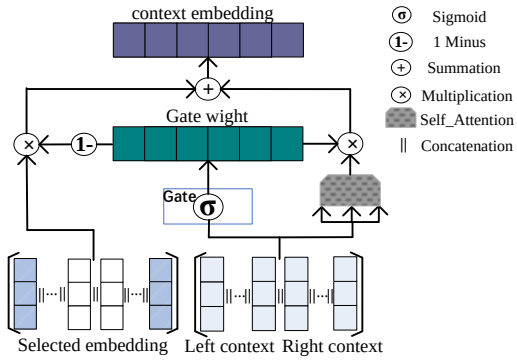

图 4 高速网络

该高速网络所对应的计算公式如式（13）~（15）所示。

$$Att\_x_w = MultiHead(x_w, x_w, x_w) \#(13)$$

$$Gate = sigmoid(W * x_w + b) \#(14)$$

$$Cont = (1 - Gate) * Select\_x_w + Gate * Att\_x_w \#(15)$$

考虑到句子中每个位置对于上下文的依赖程度不同，通过引入门控$Gate$来学习句子中每个位置所占权重，该门控机制由$sigmoid$单元组成。其中，W和b分别表示门控机制的权重矩阵与偏置项。

2.5 全局模型

由局部模型得到候选实体与该表述的上下文单词的相关性得分后，我们提出假设：每篇文章中与表述相链接的所有实体都应该具有同一个“主题”（例如前文中的“中国”指向“中国篮球队”，那么后文中的“美国”应倾向于指向“美国篮球队”），即同一篇文章内的被链接的实体应该存在制约关系，甚至会互相影响最终的链接结果。

基于上述的假设，我们的全局模型采用全连接的条件随机场（Conditional Random Field[17]，CRF），如图 1 中的 step_4 所示，定义如式（16）~（17）所示。

$$E^* = \underset{E \in C_1 \times \ldots \times C_n}{arg\,max} \sum_{i=1}^{n} \Psi(e_i, c_i) + \sum_{i \neq j} \Phi(e_i, e_j, D) \#(16)$$

$$\Phi(e_i, e_j, D) = \frac{1}{n-1} e_i^T R e_j \#(17)$$

其中，Ψ表示 2.3 节得到的局部上下文信息与候选实体的得分，Φ表示在全局模式下实体对的得分，$R \in \mathbb{R}^{d \times d}$是用来学习的对角矩阵。

Wainwright[18]等人验证等式（16）为 NP-hard 问题，受 Domke[19]等人在计算机视觉中使用类似方法的启发，以及 Le and Titov[8]等人工作的引导，允许我们通过截断消息传递进行反向传播，从而与 CRF 协同工作。本文使用最大乘积循环置信度传播（Loopy Belief Propagation[20]，LBP）为每个 mention（记为m_i）估算最大边际概率，最后再结合先验概率得到最终的$< mention - entity >$得分$\rho_i(e)$。计算如式（18）~（20）所示。

$$q(E|D) \propto exp\left\{ \sum_{i=1}^{n} \Psi(e_i, c_i) + \sum_{i \neq j} \Phi(e_i, e_j, D) \right\} \#(18)$$

$$\hat{q}_i(e_i|D) \approx \underset{e_1, \ldots, e_n (except\ e_i)}{max} q(E|D) \#(19)$$

$$\rho_i(e) = g\left(\hat{q}_i(e|D), \hat{p}(e|m_i)\right) \#(20)$$

其中，$\hat{p}(e|m_i)$表示仅限于m_i时选择实体e的概率，g为简单双层全连接神经网络，用来改变输出的维度。

3 实验设置与结果分析

本节将使用实体链接的公开数据集，通过不同的设置对模型进行实验，并对实验结果进行讨论与分析。

3.1 实验数据集

本文采用 AIDA-CoNLL[21] 数据集中的 AIDA-train 作为训练集，AIDA-A 作为验证集，AIDA-B 作为测试集，测试集还包含 Guo and Barbosa[22]发布的 MSNBC(MSB)、AQUAINT(AQ)、ACE2004(ACE) 和 WNED-WIKI(WW)，以及 Gabrilovich [23]发布的 WNED-CWEB(CWEB)。在上

述 6 个测试集中只有 AIDA-B 是和训练集属于相同的领域，其他 5 个测试集都是来自不同的领域，这又增加了实体链接的难度。表 1 详细地给出了所有语料的结构，从每篇文档中拥有的表述的个数可以看出存在一定的稀疏性问题。

表 1 实体链接数据集结构

Dataset	Number mentions	Number docs	Mentions per doc
AIDA-train	18448	946	19.5
AIDA-A	4791	216	22.1
AIDA-B	4485	231	19.4
MSB	656	20	32.8
AQ	727	50	14.5
ACE	257	36	7.1
WW	6821	320	21.3
CWEB	11154	320	34.8

本文从 2018 年的维基百科页面中抽取出 Wiki_2018 知识库，并采用半结构化的存储方式。表 2 给出了两个知识库的对比信息，由表可知 Wiki_2018 在规模上约是 Wiki_2014 知识库的 1.5 倍，蕴含更丰富的信息。

表 2 知识库对比信息

KB	Number docs	Number anchor	Size(G)
Wiki_2014	4459082	18611834	11.16
Wiki_2018	9618296	26916035	16.78

3.2 实验设置

实验中采用了 Pytorch 0.4.1 框架，并用 NVIDIA 的 1080GPU 进行加速。首先，本文实体嵌入式表示使用语料中所有候选实体以及上述 Wiki_2018 知识库进行训练，在训练过程中的得分是采用表 3 所示的 4 个指标的总和来综合衡量实体嵌入式表示的质量。

表 3 实体相关性评测结果

Metric	NDCG @1	NDCG @5	NDCG @10	MAP
(Yamada,2016) dim=500	0.59	0.56	0.59	0.52
(Ganea,2017) dim=300	0.632	0.609	0.641	**0.578**
(our,2019) dim=300	**0.635**	**0.611**	0.641	0.572

考虑到实验的可对比性，我们采用了与 Ganea[7]和 Le[8]等人相同的实验模型参数。具体的模型参数为：局部模型中表述周围的上下文单词个数K为 100，候选实体的个数S为 7，对角矩阵A和B均是由随机初始化生成，全局模型中 LBP 算法的最大循环次数为 10。优化函数选择 Adam（Adaptive Moment Estimation[24]）算法，学习率lr设置为 0.015，学习率减少步长lr_decay设置为 0.05。多头注意力机制中$head_count$设置为 50，$dropout$设置为 0.2。

3.3 实验结果及分析

实验采用准确率P、召回率R和F_1值对链接结果进行评价[25]。其中，F_1值能够综合评价模型的性能，本文使用与 Ganea[7]和 Le[8]等人相同的$Micro\text{-}F_1$。3 种评价指标的计算如式（21）~（23）所示。

$$P = \frac{Predict\_right\_num}{Predict\_num} \#(21)$$

$$R = \frac{Predict\_right\_num}{Truth\_num} \#(22)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \#(23)$$

表 4 和表 5 给出了本文和 Ganea[7]和 Le[8]模型的实验结果。可以看出，在 Wiki_2014 知识库上，我们取得了与 Le[8]相当的性能，在 AIDA-B 数据集上的性能仅比 Le[8]低了约 0.38%，其他 5 个数据集上的平均性能仅低了约 0.17%。而在规模更大的最新通用知识库 Wiki_2018 上，本文提出的 NSEL 模型能更好的与其适配，没有出现性能降低的情况。相反，在 AIDA-B 数据集上取得了目前最好的性能，同样的在其他 5 个数据集上的平均性能也比 Ganea[7]和 Le[8]的模型要高，在 MSNBC 和 WIKI 数据集上也取得了目前最好的性能。

表 4 AIDA-B 数据集实验结果

KB	Methods	AIDA-B
Wiki_2014	Ganea(2017)	92.22±0.14
	Le(2018)	**93.07±0.27**
	NSEL	92.69±0.15
Wiki_2018	Ganea(2017)	91.91±0.12
	Le(2018)	92.90±0.47
	NSEL	**93.11±0.17**

3.4 候选实体优选的效用分析

表 6 给出了 Wiki_2014 和 Wiki_2018 知识库生成的候选集中标签实体（Ground Truth）的位置分布。从表中我们可以看出规模较大的 Wiki_2018 知识库会让标签实体更多出现在候选实体集的前 7 个位置，尤其是 CWEB 语料集中多了约 26 个标签实体前移至前 7 个位置，但是也带来了在第 1 个位置实体个数减少的问题。即：增大了标签实体被链接的可能性，但降低了标签实体的初始先验概率，增加了实体消歧的难度。

表 5 其他数据集实验结果

KB	Methods	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	Avg
Wiki_2014	Ganea(2017)	93.7±0.1	88.5±0.4	88.5±0.3	77.9±0.1	77.5±0.1	85.22
	Le(2018)	**93.9±0.2**	**88.3±0.6**	**89.9±0.8**	**77.5±0.1**	**78.0±0.1**	**85.51**
	NSEL	94.0±0.1	88.1±0.5	89.1±0.5	77.6±0.1	77.9±0.1	85.34
Wiki_2018	Ganea(2017)	93.8±0.1	87.0±0.5	87.8±0.4	77.5±0.1	77.6±0.1	84.74
	Le(2018)	94.2±0.2	87.5±0.8	88.7±0.6	77.8±0.1	77.5±0.1	85.14
	NSEL	**94.2±0.1**	**88.1±0.6**	**89.3±0.3**	**77.9±0.1**	**77.9±0.1**	**85.48**

表 6 标签实体在候选集中的位置分布

Corpus	KB	Position of Ground Truth in Entity Candidate Set								Top 7
		1	2	3	4	5	6	7	≥8	proportion
AIDA-B	Wiki_2014	3084	650	184	108	116	37	16	290	93.53
	Wiki_2018	3082	656	172	126	105	35	18	291	93.51
MSNBC	Wiki_2014	496	70	32	6	2	0	3	47	92.83
	Wiki_2018	488	78	32	7	2	0	3	46	92.98
AQUAINT	Wiki_2014	604	56	10	5	4	0	1	47	93.53
	Wiki_2018	605	55	11	6	3	0	1	46	93.67
ACE2004	Wiki_2014	217	5	4	1	0	1	0	29	88.71
	Wiki_2018	209	13	4	1	0	1	0	29	88.71
CWEB	Wiki_2014	7430	1323	474	254	103	86	67	1417	87.29
	Wiki_2018	7428	1318	489	275	98	90	65	1391	87.52
WIKI	Wiki_2014	4377	878	306	154	115	70	71	850	87.53
	Wiki_2018	4367	866	320	149	107	74	78	860	87.39

表 7 详细实验对比结果

Methods	KB	Corpus	P	R	F_1
Top 7 Selection	Wiki_2018	AIDA-B	91.97	91.97	91.97
Optimized Selection	**Wiki_2018**	**AIDA-B**	**93.11**	**93.11**	**93.11**
Top 7 Selection	Wiki_2018	MSNBC	93.5	92.3	92.9
Optimized Selection	**Wiki_2018**	**MSNBC**	**94.6**	**93.8**	**94.2**
Top 7 Selection	Wiki_2018	AQUAINT	88.0	85.7	86.8
Optimized Selection	**Wiki_2018**	**AQUAINT**	**89.4**	**86.8**	**88.1**
Top 7 Selection	Wiki_2018	ACE2004	88.6	87.8	88.2
Optimized Selection	**Wiki_2018**	**ACE2004**	**89.7**	**88.9**	**89.3**
Top 7 Selection	Wiki_2018	CWEB	77.0	76.8	76.9
Optimized Selection	**Wiki_2018**	**CWEB**	**78.1**	**77.7**	**77.9**
Top 7 Selection	Wiki_2018	WIKI	77.3	76.9	77.1
Optimized Selection	**Wiki_2018**	**WIKI**	**78.0**	**77.8**	**77.9**

我们将采用前 7 个候选实体和 2.2 节候选实体优选的策略在 6 个数据集上进行对比实验，实验结果如表 7 所示。在 6 个公开数据集上的实验结果显示，采用候选实体优选策略的性能在P、R和F_1值上

均高于采用前 7 个候选实体策略。即：本文提出的融合上下文相关度和先验概率的候选实体优选策略能很好的解决规模较大的知识库对候选实体的先验概率的负面影响，同时能很好的利用其带来的标签实体更高覆盖率的正面影响。

3.4 Self-attention 的效用分析

最后，我们进一步比较未引入 Self-attention 的 NEL 模型和引入 Self-attention 的 NSEL 模型在 AIDA-B 数据集上的实验结果，分析 Self-attention 的贡献度。表 8 给出了使用 Wiki_2018 知识库的实体链接性能，从中可以看出，NSEL 模型在 6 个数据集上的各项性能均高于 NEL 模型，验证了引入 Self-attention 的效用。模型可以从多个不同子空间捕获上下文信息，更好地理解句子结构，学习到了更丰富的上下文表征，从而P、R和F_1值均有所提升。

表 8 详细实验对比结果

Methods	Corpus	P	R	F_1
NEL	AIDA-B	92.26	92.26	92.26
NSEL	**AIDA-B**	**93.11**	**93.11**	**93.11**
NEL	MSNBC	93.7	93.1	93.4
NSEL	**MSNBC**	**94.6**	**93.8**	**94.2**
NEL	AQUAINT	88.6	86.2	87.4
NSEL	**AQUAINT**	**89.4**	**86.8**	**88.1**
NEL	ACE2004	88.6	88.2	88.4
NSEL	**ACE2004**	**89.7**	**88.9**	**89.3**
NEL	CWEB	77.3	77.1	77.2
NSEL	**CWEB**	**78.1**	**77.7**	**77.9**
NEL	WIKI	77.4	77.4	77.4
NSEL	**WIKI**	**78.0**	**77.8**	**77.9**

4 结论

本文提出了一个融合局部和全局信息的自注意力机制与高速网络相结合的神经网络实体链接框架，该方法很好地使用了最新通用知识库所蕴含的丰富信息。在 6 个实体链接公开数据集上的实验结果也表明了该方法的有效性，在 Wiki_2018 通用知识库上取得了目前最好的性能。

未来我们将考虑使用双语言或多语言的知识库进行联合学习，利用不同语言之间的互补性进一步提升实体链接的性能。

参考文献

[1] Yih S W, Chang M W, He X, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[J]. 2015.

[2] Ji H, Nothman J, Hachey B, et al. Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking[C]//TAC. 2015.

[3] Ji H, Nothman J, Dang H T, et al. Overview of TAC-KBP2016 Tri-lingual EDL and its impact on end-to-end Cold-Start KBP[J]. Proceedings of TAC, 2016.

[4] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 541-550.

[5] Sun Y, Lin L, Tang D, et al. Modeling mention, context and entity with neural networks for entity disambiguation[C]//Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.

[6] Francis-Landau M, Durrett G, Klein D. Capturing semantic similarity for entity linking with convolutional neural networks[J]. arXiv preprint arXiv:1604.00734, 2016.

[7] Ganea, Octavian-Eugen, and Thomas Hofmann. "Deep joint entity disambiguation with local neural attention." arXiv preprint arXiv:1704.04920 (2017).

[8] Le, Phong, and Ivan Titov. "Improving entity linking by modeling latent relations between mentions." arXiv preprint arXiv:1804.10637 (2018).

[9] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[10] Ceccarelli D, Lucchese C, Orlando S, et al. Learning relatedness measures for entity linking[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013: 139-148.

[11] Busa-Fekete R, Szarvas G, Elteto T, et al. An apple-to-apple comparison of Learning-to-rank algorithms in terms of Normalized Discounted Cumulative Gain[C]//20th European Conference on Artificial Intelligence (ECAI 2012): Preference Learning: Problems and Applications in AI Workshop. Ios Press, 2012, 242.

[12] Yue Y, Finley T, Radlinski F, et al. A support vector method for optimizing average precision[C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 271-278.

[13] Spitkovsky V I, Chang A X. A cross-lingual dictionary for english wikipedia concepts[J]. 2012.

[14] Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 782-792.

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 6000-6010.

[16] Srivastava R K, Greff K, Schmidhuber J. Highway

networks[J]. arXiv preprint arXiv:1505.00387, 2015.
[17] 洪铭材, 张阔, 唐杰,等. 基于条件随机场(CRFs)的中文词性标注方法[J]. 计算机科学, 2006, 33(10):148-151.
[18] Wainwright, Martin J., and Michael I. Jordan. "Graphical models, exponential families, and variational inference." Foundations and Trends® in Machine Learning 1.1–2 (2008): 1-305.
[19] Denton, Emily, et al. "User conditional hashtag prediction for images." Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2015.
[20] Murphy, Kevin P., Yair Weiss, and Michael I. Jordan. "Loopy belief propagation for approximate inference: An empirical study." Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999.
[21] Hoffart, Johannes, et al. "Robust disambiguation of named entities in text." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.
[22] Guo, Zhaochen, and Denilson Barbosa. "Robust named entity disambiguation with random walks." Semantic Web 9.4 (2018): 459-479.
[23] Gabrilovich, Evgeniy, Michael Ringgaard, and Amarnag Subramanya. "FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0)." Note:http://lemurproject.org/clueweb09/FACC1/Cited by 5 (2013).
[24] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
[25] Chinchor N. MUC-4 evaluation metrics[C]//Proceedings of the 4th conference on Message understanding. Association for Computational Linguistics, 1992: 22-29.

李明扬（1995—），硕士研究生，主要研究领域为自然语言处理、实体链接、知识图谱。
E-mail：20175227067@stu.suda.edu.cn

姜嘉伟（1997—），本科生，主要研究领域为自然语言处理、知识图谱。
E-mail：jiang_jiawei@outlook.com

孔芳（1977—），博士，主要研究领域为机器学习、自然语言处理、篇章分析。
E-mail：kongfang@suda.edu.cn