知识增强的实体消歧与实体类别判断

潘春光 [1] 王胜广 [1] 罗志鹏 [1]

[1] 深兰科技（上海）有限公司

摘要 论文基于 CCKS2020：面向中文短文本的实体链指任务给定的数据集，基于该数据集任务可分为多歧义实体消歧和 NIL 实体的上位概念类型判断两个子任务。对于实体消歧任务提出了基于 BERT 和实体特征的实体消歧模型，模型对每一个候选实体进行预测，然后对预测的概率进行排序，由于数据集中包含 NIL 实体，模型将 NIL 实体作为候选实体参与模型训练和概率排序，进而完成消歧任务。针对 NIL 实体的上位概念类型判断任务提出了基于问答的 NIL 实体类型判断模型，模型通过构建问句并依据已知实体信息构建上下文，有效的引入短文本中已知实体的知识库信息，额外信息的引入能够显著提升了模型的性能。本文提出的方案获得了 A 榜第一名（F1 为 0.89193），B 榜第二名（F1 为 0.89538）的成绩。

Keywords: 实体链接 实体消歧 实体类型推断

1 引言

如今的时代是互联网的时代每天都产生了大量的网络数据，所以如何从大量数据中提取有用的知识信息成为现在研究的热门话题，而实体链接在这个过程中发挥着关键的作用[1][2]。网络上的数据大多是以网络文本的形式呈现的，这些网络文本中存在大量命名实体[3]，它们是人们理解网络文本的基本元素。但是当这些实体出现在不同情境，它的意思也变得非常模糊，因为一个命名实体有多个名称，同样一个名称也可以表示几个不同的命名实体。因此需要将这些识别到的实体链接到现有的知识库中，这样人们才能够更加准确地理解实体所指的意义。另一方面，随着谷歌知识图谱的发展以及维基百科等知识共现社区的出现，自动化构建知识（Wikipedia[4] 、 DBpedia[5]、YAGO[6]等）变得越来越重要。自动化构建知识库需要从网络文本中提取实体与实体之间的关系，并将其添加到知识库中。而在添加到知识库之前最重要的一步是需要消除文本中识别的实体与知识库中实体之间的歧义，实体链接在这个过程中发挥着至关重要的作用。

实体链接是一项识别文本中的实体指称（指文本被识别到的命名实体[7]）并将其链接到知识库中对应实体上的任务[1]。对于一个给定的实体链接任务，

首先需要使用命名实体识别方法和工具识别文本中的实体，然后对每个实体指称利用候选实体生成技术生成对应候选实体集，最后利用文本信息和知识库的信息消除候选实体的歧义，通过每个候选实体的得分选取最高的作为匹配实体，如果最高的得分小于某个阈值则将实体指称标记为 NIL（代表没有对应实体）。一般来讲，实体链接包括三个主要环节：命名实体识别、候选实体生成、候选实体消歧。

由于实体具有的意思具有高度模糊性，除了有一词多义的情况，另一个常见的情况是多词同义。在日常生活中，人们经常会用不同的名字去指代一个人，例如阿里巴巴马云，人们经常使用的称呼有：马云、马爸爸、风清扬等。这两种情况是造成实体链接任务困难的主要因素。而相比较英文以及长文本，针对中文短文本的实体链接存在更大的挑战。主要有以下几个难点：

（1）相比较英文单词中间由空格隔开，中文文本由字紧密排列组成，容易造成交叉歧义和组合歧义；

（2）短文本上下文语境不丰富，相对于长文本，短文本的语境理解更加困难；

（3）短文本口语化严重容易造成表达不够完整。

因此，不同于英文文本或者长文本，中文文本由于其文本的复杂、信息的缺失导致其实体链接任务效果较差，只有解决上述几个难点才能有效提升当前短文本实体链接的效果。

本文基于“CCKS 2020: 面向中文短文本的实体链指任务”，对比 2019 年任务去掉了实体识别，专注于中文短文本场景下的多歧义实体消歧技术，增加对新实体（NIL 实体）的上位概念类型判断同时对标注文本数据调整，增加多模任务场景下的文本源，同时调整了多歧义实体比例。为此，基于数据集任务可分为多歧义实体消歧和 NIL 实体的上位概念类型判断两个子任务，针对实体消歧子任务，提出了基于 BERT[8]和实体特征的实体消歧模型，针对 NIL 实体的上位概念类型判断提出了基于问答的 NIL 实体类型判断模型。

2 实体消歧

实体消歧是实体链接最关键的一步，同时作为自然语言处理的一项基本任务，它在搜索引擎、问答系统、对话系统等应用中都扮演着重要的角色。实体消歧主要是对于给定的实体指称，利用候选实体生成技术得到相应的候选实体集，然后在利用实体消歧模型找到真正对应的那个实体。

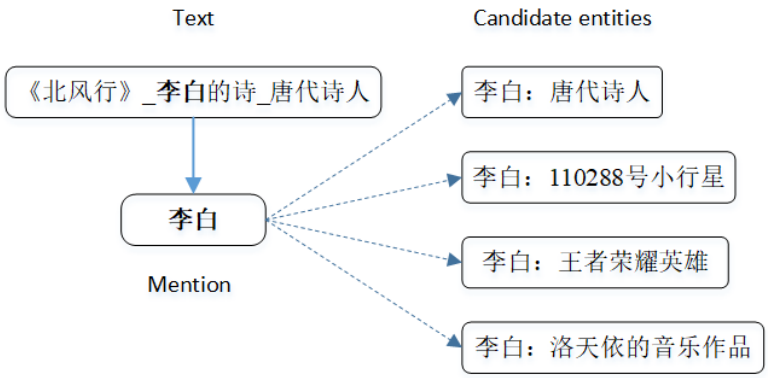

图 1 候选实体生成

如图 1 所示，利用短文本中的实体指称，可以通过候选实体生成的方式得到候选实体集合，而实体消歧的目的就是在这部分候选实体集合中找到对应的那个实体，如果找不到，则用 NIL（代表没有对应实体）表示。具体过程如下：

给定短文本输入（用 Query 表示），此 Query 中有 *N* 个实体 mention，如式(1)所示：

$$M_q = \{m_1, m_2, m_3, \ldots\} \tag{1}$$

每个实体 mention 链接到知识库的实体 id 为，如式(2)所示：

$$E_q = \{e_1, e_2, e_3 \ldots\} \tag{2}$$

则实体消歧可以抽象为如式(3)所示：

$$E_q = \underset{\mathrm{E}}{\operatorname{argmax}} \sum_{i=1}^{N} \emptyset(m_i, e_i) \tag{3}$$

其中E代表实体指称m_i的一个候选实体集合，其中e_i为实体指称m_i对应的第 i 个实体。$\emptyset(m_i, e_i)$是计算指称m_i和实体e_i之间的匹配程度。最后选取匹配程度得分最高的作为对应实体，如果得分最高的实体，分数小于某个阈值，则认为该实体指称没有对应实体。从以上介绍可以看出，实体消歧过程主要包括两个步骤，即候选实体生成和实体消歧。由于短文本大多来自微博评论、文本搜索、日常对话，往往存在以下几个问题。例如，文本内容较短、口语化严重、实体通常用缩写、简称、别名等进行表示。因此短文本实体消歧具有更大的挑战。针对候选实体生成，本文采用基于字典的方式，对于实体消歧为了提取更深的语义特征，提高模型消歧的能力，本文采用预训练模型 BERT 作为基础网络结构，提出了基于 BERT 和实体特征的实体消歧模型。

2.1 候选实体生成

候选实体生成是指利用命名实体识别方法识别出文本中的实体指称，然后依据实体名字并结合其他特征在知识库中找到对应的候选实体。候选实体生成主要有以下两种方法：基于字典的方式[9] [10] [11] [12]和基于搜索引擎的方式[13] [14]，本文结合数据特点采用了基于字典的方式。

候选实体生成最方便的方法就是基于字典的方法，这种方法需要根据给定的知识库构建名称字典，字典的键就是实体的名字，而值则是这个名字所对应的所有的具有相同名字的实体。例如对于实体名字“苹果”，可能的实体包括：苹果（苹果公司）、苹果（蔷薇科苹果亚科苹果属植物）、苹果（李玉执导电影）。对于每个实体指称，该方法都要检索字典的键，如果字典的键符合要求，则将该键对应的值中所有的实体都加入到候选实体集中。其中判断字典的键是否符合要求通常的做法是采用精确匹配的方式，只有当实体指称和字典键彼此完全匹配的情况下才加入到候选实体集，本文采用为基于字典的精确匹配方式。基本流程为先通过知识库中的实体名字以及实体别称构建实体字典，然后采用精确匹配的方式匹配得到候选实体。

2.2 实体消歧模型

实体消歧现在较为流行的方法有两类：基于二分类方法[15] [16]和基于 Rank 的方法[17]。利用基于二分类方法和基于 Rank 的方法能够得到实体指称与每个候选实体的匹配得分，最后选取最高的作为匹配实体，实体消歧需要计算指称和实体之间的匹配程度，本文采用二分类的思想来计算匹配程度。基于二分类模型进行实体消歧，基本步骤为：

(1) 构建数据集。
(2) 构建二分类模型。
(3) 候选实体预测得分排序，得到最终结果。

图 2 实体消歧样例

如图 2 所示，每个短文本有多个实体指称，针对其中一个实体指称能够得到多个候选实体，在多个候选实体中，最多只有一个为正确的实体，也可能不存在。基于二分类的实体消歧的详细步骤如下：

(1) 构建数据集

由于给定的数据为实体指称、以及实体指称链接到知识库的实体对应的id，由于没有负样本，数据不能直接用在二分类的消歧模型上。本文将短文本的实体指称和其对应的知识库的实体 id 构成一个正样本，然后利用候选实体生成方法，生成对应的候选实体集，在候选实体集中选取负样本。由于加入了NIL 实体，本文将 NIL 实体也作为候选实体参与训练和排序，当实体指称有正样本的时候 NIL 实体为负样本，当实体指称没有正样本的时候 NIL 实体为正样本。

(2) 构建二分类模型

模型的输入为短文本、实体指称、以及候选实体的描述信息。二分类模型将正样本的 label 视为 1，负样本视为 0。一般利用实体指称所在上下文的特征，以及候选实体的描述文本的特征进行分类。模型最后会输入一个概率值，概率值越接近 1，代表匹配的程度越高。

(3) 候选实体预测得分排序，得到最终结果

针对短文本中的每个实体指称循环与所有的候选实体进行二分类，得到概率得分，将当前实体指称与所有候选实体的得分排序，选取最高的作为正确实体，如果最高分的候选实体为 NIL 实体则认为该实体为 NIL 实体，如果最高的得分小于给定的阈值，也认为该实体为 NIL 实体。

2.3 模型结构

现在比较流行的消歧模型常用的方案是提取实体所在短文本的上下文特征，以及候选实体描述文本的特征，在将这两类特征经过全连接网络，最后进行二分类。这类方法对于长文本很有效，但是对于短文本的消歧效果不是很好，主要是因为短文本内容较短，上下文太少，难以提取有效的上下文特征，而候选实体的描述文本过长，这种情况造成传统的实体消歧模型效果并不理想。本文考虑到短文本的特性，提出了基于 BERT 和实体特征的消歧模型。

模型如图 3 所示，模型采用的思想主要是为利用 BERT 模型[CLS]符号的输出向量，以及实体指称所在位置的特征向量，经过全连接层，然后经过 sigmoid 进行二分类。其中[CLS]符号的输出向量可以用来判断短文本和候选实体的描述文本是否处在同一语义场景，实体位置的特征向量可以代表实体的上下文特征。实体的上下文特征主要为了解决实体名字相同但是链接实体不同的情况，如：“海绵宝宝：抠门儿的蟹老板为了省 5 块钱，要辞退海绵宝宝？”，前面海绵宝宝代表动画片名字，后面海绵宝宝代表剧中角色。

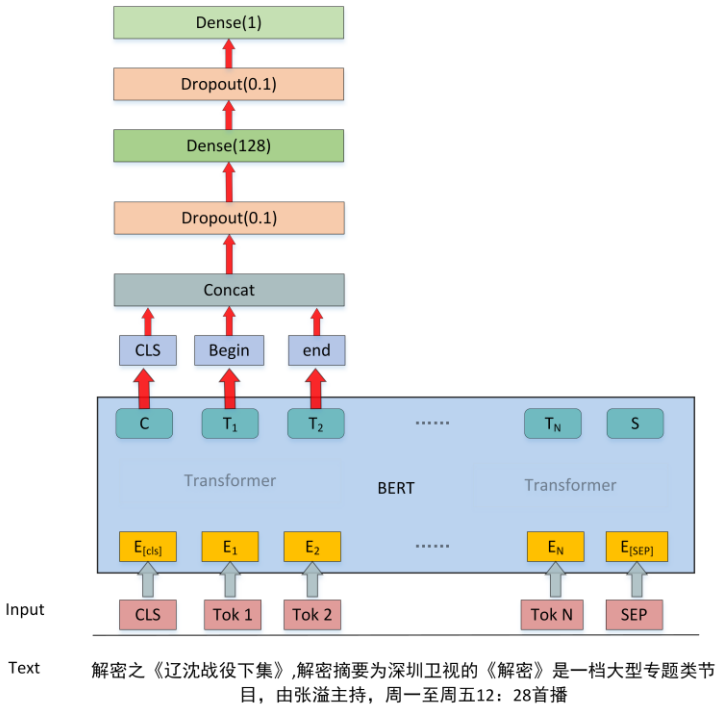

图 3 实体消歧模型图

模型的输入为短文本以及候选实体的描述文本，形式为：[CLS]短文本[SEP]候选实体描述文本[SEP]。不同于以往在训练前选取固定的负样本，模型采用动态负采样技术，在模型训练中每个 epoch 选取不同的负样本参与训练，通过这种方式能够极大的提高模型的泛化能力，由于增加了 NIL 实体，对 NIL 实体也作为候选实体参与训练和排序。

3 实体类型判断

当文本中的实体指称在现有知识库中不存在可链接的实体时，此实体指称被称作 NIL 实体，任务要求预测出 NIL 实体对应的类型。对于实体类型判断任务一般使用基于神经网络的方法，当分类问题来做，一个基本的思路也就是 baseline 思路为，通过提取实体指称位置的向量，然后经过全连接分类，得到实体的类型。这种方案的最大缺点就是仅仅用短文本的信息去对 NIL 实体进行类型分类，没有利用到已知实体信息的特征，为了利用上其他不是 NIL 实体的信息，本文构建了基于问答的实体类型判断模型，模型基于问答的思想，通过构建问句和上下文将已知实体的信息输入到模型中，来提升实体类别判断的性能。

例如对于句子，“神探加杰特，和彭妮长得一模一样，竟想要霸占泰龙的位置” 其中实体有“神探加杰特” 、“彭妮”，“泰龙”，“位置” 4 个实体，需要预测类型的 NIL 实体有“彭妮”，“泰龙”，对于实体“彭妮”，“泰龙”，“彭妮” 根据命名习惯很容易判断为类型为 Person 类型， 就算根据短文本的语义分析判断“彭妮”依旧是 Person 类型，同样模型学习到的也是 Person 类型。而“彭妮”的真实类型为 VirtualThings 类型，可以看出在没有其他额外信息的情况下，很难准确预测“彭妮”的类型。任务要求仅仅预测 NIL 实体的类型即可，其他实体的类型我们通过实体消歧已经知道了。那么实体类型判断任务可以变成，在已知部分实体信息的情况下，求短文本中其他实体的类型。对于这种任务我们采用问答的形式，对于上述例子根据短文本以及 mention 构建问句为：、

“神探加杰特，和彭妮长得一模一样，竟想要霸占泰龙的位置，彭妮的类型是什么？”

根据已知的实体信息构建相关上下文为：

“神探加杰特的类型是作品，描述为 1983--1985 年美国播出的动画片，位置类型为其他，描述为词语释义”

通过上下文信息，可以得知“神探加杰特”为一个动画片，而不是电视剧电影等其他作品，那么模型则可以轻易学出，动画片中的人物为 VirtualThings 类型。本文将已知实体的信息分为实体类型和实体的描述，后续会对此进行实验分析。

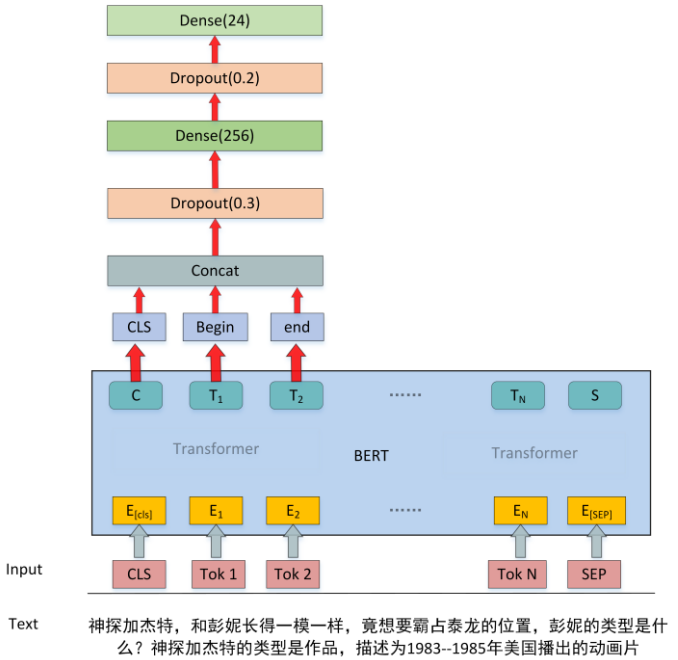

图 4 实体类别判断模型

模型基于预训练模型 BERT，模型如图 4 所示，模型利用 BERT 模型[CLS]符号的输出向量，以及实体所在位置的特征向量，经过全连接层，然后经过 Softmax 激活函数进行多分类。模型的输为：[CLS]短文本[SEP]已知信息的上下文文本[SEP]。为了提高模型的泛化能力，在训练过程中本文加入了对抗训练，其中采用的对抗训练方法为 FGM，具体步骤如下：

对于每个输入 x:

1.计算 x 的前向 loss、反向传播得到梯度

2.根据 embedding 矩阵的梯度计算出 r，并加到当前 embedding 上，相当于 x+r

3.计算 x+r 的前向 loss，反向传播得到对抗的梯度，累加到(1)的梯度上

4.将 embedding 恢复为(1)时的值

5.根据(3)的梯度对参数进行更新

4 结果分析

数据集由百度 CCKS2020 提供，数据分为 7 万训练集和 1 万验证集，后续实验结果均为验证集的实验结果。实体链接评价方式采用官方提供的评价方式，具体方式为，通过将输出结果与人工标注的集合进行比较来计算准确率(Precision)，召回率(Recall)和 F-1 分值(F-1 score)。 具体计算过程如下所示：

给定短文本输入（用 Text 表示，其属于 golden 标注集），此 Text 中有 N 个 mention：$M_n = \{m_1, m_2, m_3 \dots m_n\}$，每个 mention 链接到知识库的实体 id 为：$E_n = \{e_1, e_2, e_3 \dots e_n\}$，实体标注系统输出标注结果如下：$E'_n = \{e'_1, e'_2, e'_3 \dots e'_n\}$，则实体标注的准确率定义如下：

$$\mathrm{P} = \frac{\sum_{n \epsilon N} |E_n \cap E'_n|}{\sum_{n \epsilon N} |E'_n|}$$

实体标注的召回率定义如下：

$$\mathrm{R} = \frac{\sum_{n \epsilon N} |E_n \cap E'_n|}{\sum_{n \epsilon N} |E_n|}$$

实体标注的 F1 值定义如下：

$$F1 = \frac{2 * P * R}{P + R}$$

在计算评价指标时，对 NIL 实体的上位概念类型判断结果 NIL_Type 与实体的关联 id 等价处理。

针对实体消歧子任务，采用和实体链接一样的评价方式，只不过不对 NIL 实体做类别判断，对于实体类别判断子任务采用准确率作为评价标准。

4.1 实体消歧实验

在实体消歧实验阶段，本文对 NIL 实体类型不做识别，下述试验结果没有加入 NIL 实体的类型判断。相关实验中参数配置如下：batch size 为 32，针对不同层采用不同的学习率，其中 BERT 模型初始学习率为 1e-5，其他模型参数初始学习率为 5e-4，为了更好地收敛到最优，采用了基于指数衰减的学习率衰减策略，每轮衰减为原来的 0.5 倍。

NIL 实体判定策略，针对某个实体指称，有三种情况会判定为 NIL 实体，1. 没有候选实体 2. 候选实体得分排序，最高得分是 NIL，3. 候选实体得分排序，最高得分不是 NIL，但是得分小于 0.05。

表 1 实体消歧实验结果表.

编号	模型	F1
1	Model-static- neg2-ernie	0.7376
2	Model-dyanmic- neg2-ernie	0.7415
3	Model-dyanmic- neg1-ernie	0.7402
4	Model-dyanmic- neg3-ernie	0.7393
5	Model-dyanmic- neg2-bert	0.7380
6	Model-dyanmic- neg2-roberta	0.7383

本文对以下几个维度进行了实验分析，1.动态负采样与静态负采样 2.负采样的个数 3.不同的预训练模型。实验结果如表 1 所示，static 代表静态负采样，dyanmic 代表动态负采样，neg2 代表负采样的个数为 2。相关实验结果如表 1 所示。

对比模型 1 与模型 2，动态负采样对比静态负采样有着巨大的提升，静态负采样之所以不好是没有利用上更多负样本的信息，对比模型 2、3、4 可以得出，负样本个数并不是越多越好，当负样本太多时会造成类别不均衡导致最终性能下降，并且负样本增多也会导致数据变大，训练时间增加，综合考虑本文最终采用 2 个负样本。对比模型 4、5、6 可以看出不同的预训练模型对最终的结果也有着很大的影响，由于数据集由百度提供，数据中知识库来源大多来自百度百科，所以百度开源的 ernie 模型性能最佳，roberta 次之，综合考虑最终本文采取的方案为 ernie 模型与 roberta-wwt 融合。

4.1 实体类型判断实验

对于实体类型判断任务，实验参数与实体消歧相同，在实验设计方向本文主要设计了以下几个方面的对照实验，模型 1: 采用 baseline 思路，将实体位置的向量特征输出分类，模型 2：仅仅加入已知实体的类型信息，不加入其他信息，模型 3：加入简短的实体描述信息，模型 4：既加入实体类型信息，又加入

实体描述信息，模型 5：在模型 4 的基础上加入对抗训练，相关实验结果如表 2 所示。

表 2 实体类型判断实验结果表

编号	模型	准确率
1	Model-baseline	0.8628
2	Model-type	0.8793
3	Model-desc	0.8782
4	Model-type+desc	0.8800
5	Model- type+desc+FGM	0.8819

由模型 1、2、3、4 可以发现，相比于 baseline 加入已知实体的信息能够显著提升模型的性能，说明已知实体的信息对短文本 NIL 实体的类别判断有着很大的帮助。对比模型 2 和模型 3 可以发现，已知实体的类别信息更为重要，这可能与仅仅使用了简短的描述信息而没有使用实体的全部描述信息有关，从模型 4 可以得出将实体的类型信息与描述信息相结合则能够达到更好的效果。从模型 5 可以看出加入对抗训练能够提升模型的泛化能力，对最终的结果也有很大的提升。

表 3 实体链接结果表.

编号	模型	F1
1	开发集	0.88010
2	测试集 A	0.89193
3	测试集 B	0.89538

最终将两个模型合在一起得到最终的实体链接效果表如表 3 所示，其中测试集 A 和测试集 B 为多个模型交叉验证求平均的结果。

5 总结

本文提出的模型在 CCKS2020（全国知识图谱与语义计算大会）举办的“面向中文短文本的实体链指”评测任务中取得了 A 榜第一名的成绩，B 榜第二名的成绩，其中基于问答的实体类型判断模型巧妙的利用到了已知实体的信息，此模型对后续此类型的任务提供了借鉴意义，但是本文依旧有些地方需要改进，如知识库给定的数据具有一定的结构信息，直接连在一起当文本使用，可能会损失结构信息特征，可以设计更好的网络来利用这部分结构化信息。

参考文献

1. Rao, Delip, Paul McNamee, and Mark Dredze. Entity linking: Finding extracted entities in a knowledge base[J]. Multi-source, multilingual information extraction and summarization. Springer, Berlin, Heidelberg, 2013. 93-115.
2. L. Derczynskiet al, Analysis of named entity recognition and linking for tweets[J]. Information Processing & Management, 2015, 51.2: 32-49.
3. Sundheim, B. M. (1995). Named entity task definition, version 2.1[C]. In Proc. Sixth Message Understanding Conf. (MUC-6), Nov. 1995,317-332.
4. GB/T 7714Melanie, Remy. Wikipedia: The Free Encyclopedia [J]. Reference Reviews, 1997.
5. Lehmann J, Isele R, Jakob M, et al. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia[J]. Semantic Web, 2014, 6(2).
6. Rebele, Thomas, Suchanek, Fabian M, Hoffart, Johannes,等. YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames[M]. The Semantic Web – ISWC 2016. Springer International Publishing, 2016.
7. Grishman R. Message understanding conference-6 : A brief history[C]. Proceedings of the 16th conference on Computational linguistics. 1996.
8. Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
9. W. Shen, J. Wang, P. Luo, and M. Wang. Linden: Linking named entitieswith knowledge base via semantic knowledge[C]. inProc. WWW, 2012,pp. 449–458.
10. Zhang W, Su J, Tan C L, et al. Entity Linking Leveraging Automatically Generated Annotation[C]. COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China. DBLP, 2010.
11. J. G. Zheng, et al, Entity linking for biomedical literature[C]. BMC medical informatics and decision making, 2015, 15.S1: S4.
12. Zhang W, Sim Y C, Su J, et al. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling[C]. IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011. AAAI Press, 2011.
13. M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin. Entity disambiguation for knowledge base population[C]. inCOLING, 2010, 277–285.
14. X. Han and J. Zhao. Nlprkbp in tac 2009 kbp track: A two-stage method to entity linking[C]. inTAC 2009 Workshop, 2009.
15. Chen Z, Ji H. Collaborative Ranking: A Case Study on Entity Linking[C]. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP

2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL. Association for Computational Linguistics, 2011.

16. Pilz A, Gerhard Paaß. From names to entities using thematic context distance[C]. Acm Conference on Information & Knowledge Management. ACM, 2011.
17. R.C.Bunescu andM.Pasca. Using encyclopedic knowledge for named entity disambiguation[C]. in Proc. EACL, 2006, pp. 9–16.