文章编号：1003-0077(2016)04-0176-08

融合多种特征的实体链接技术研究

陈玉博[1]，何世柱[1]，刘 康[1]，赵 军[1]，吕学强[2]

(1. 中国科学院自动化研究所，模式识别国家重点实验室，北京 100190；
2. 北京信息科技大学 网络文化与数字传播北京市重点实验室，北京 100101)

摘 要：实体消歧是自然语言理解的重要研究内容，旨在解决文本信息中普遍存在的命名实体歧义问题，在信息抽取、知识工程和语义网络等领域有广泛的应用价值。实体链接是实体消歧的一种重要方法，该方法将具有歧义的实体指称项链接到给定的知识库中从而实现实体歧义的消除[1]。传统的实体链接方法主要利用上下文的词语匹配等表层特征，缺乏深层语义信息，针对这一问题，该文提出的实体链接方法利用了多种特征，从不同的维度捕获语义信息。为了更好地融合各个维度的特征，该文利用了基于排序学习框架的实体链接方法，与传统的方法相比，节省了人工对大量的模型参数选择和调节的工作，与基于分类的方法相比，能更好地利用到候选之间的关系信息。在 TAC-KBP-2009 的实体链接评测数据上的实验表明，该文提出的特征和方法表现出良好的性能，在评测指标上高出参赛队伍最好水平 2.21%，达到 84.38%。

关键词：实体消歧；实体链接；排序学习

中图分类号：TP391 **文献标识码**：A

Entity Linking Based on Multiple Features

CHEN Yubo[1], HE Shizhu[1], LIU Kang[1], ZHAO Jun[1], LV Xueqiang[2]

(1. NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;
2. ICDDR, Beijing Information Science and Technology University, Beijing 100101, China)

Abstract: Entity linking is an important method of entity disambiguation, which aims to map an entity to an entry stored in the existing knowledge base. Several methods have been proposed to tackle this problem, most of which are based on the co-occurrence statistics without capture various semantic relations. In this paper, we make use of multiple features and propose a learning to rank algorithm for entity linking. It effectively utilizes the relationship information among the candidates and save a lot of time and effort. The experiment results on the TAC KBP 2009 dataset demonstrate the effectiveness of our proposed features and framework by an accuracy of 84.38%, exceeding the best result of the TAC KBP 2009 by 2.21%.

Key words: Named Entity disambiguation; entity linking; learning to rank

1 引言

近年来，随着互联网的普及和迅速发展，越来越多的信息以数字化的方式存储在网络中。如何在浩繁的数据中实现深层语义检索和查询已经引起了众多学者的关注。为了实现这一目标，必须构建出机器可以理解的、组织良好的结构化知识库或知识图谱。目前已经有很多公开的结构化知识库，例如，YAGO[2]、KOG[3]和 DBpedia[4]等。在构建和维护结构化知识库时，不可避免地会遇到命名实体歧义的问题。因此，研究实体链接技术具有重要的学术价值和现实意义。

命名实体歧义指的是同一个实体指称项在不同的上下文中可以对应到不同真实世界实体的语言现象。例如，给定如下两个包含“Michael Jordan”的句子：

• Michael Jordan is a famous american bas-

收稿日期：2014-09-15 定稿日期：2015-03-20

基金项目：国家自然科学基金(61202329, 61272332)；网络文化与数字传播北京市重点实验室开放课题(ICDD201201)

ketball player.

• Michael Jordan is a famous professor in the field of machine learning.

上述例子中的两个“Michael Jordan”分别对应着篮球运动员“Jordan”和机器学习领域的教授“Jordan”。实体链接系统的主要任务是将文本中具有歧义的实体指称项链接到知识库中的相应实体上，如果在知识库中没有相对应的实体，则链接到空实体上。实体链接中的关键问题是候选实体与实体指称项间的语义相似度的计算。传统的研究工作中主要利用词袋子模型计算指称项所在上下文文本与候选实体所在文本之间的文本相似度，进而用文本的相似度来衡量实体间的相似度，还有学者将类似的表层语义信息作为主要特征来判断实体间的相似度。但是类似的表层语义特征都是基于词匹配的，缺乏深层语义信息。不适用同一实体出现的上下文语境没有匹配词汇，或者匹配词汇数量少的情况。例如，我们假设知识库中有两个名为“Michael Jordan”的实体：

• 实体名：Michael Jordan(NBA Player)
  文本：Michael Jordan plays basketball in Chicago Bulls.

• 实体名：Michael Jordan(Machine Learning Professor)
  文本：Michael Jordan is a famous professor in the field of machine learning.

当待消歧的实体指称项“Michael Jordan”出现在文本“Michael Jordan wins NBA MVP.”中时，实体指称项应当链接到美国篮球运动员迈克尔乔丹上，因为消歧文本中的“MVP”和知识库实体“Michael Jordan”定义中的“basketball”和“Chicago Bulls”有非常高的语义关联度。但上述例子中，除了实体名外，实体指称项所在的文本与知识库中该实体的描述文本没有匹配的词，导致传统基于词袋的模型无法取得满意的结果。为了解决这一问题，本文挖掘并利用 Wikipedia 中的实体关联知识，提出了一系列包含深层语义信息的特征，并将这些深层语义特征与表层字面特征融合，完成实体链接。为了更好地利用本文所提出的特征信息，在消歧阶段本文利用了基于排序学习框架。相较于基于分类的实体消歧方法，基于排序学习的方法能更好地考虑候选实体间的关系。本文设计并实现了一个完整的实体消歧系统，系统由候选实体生成模块和候选实体选择模块两部分组成。进行实体消歧任务时，主要分两个步骤完成：(1)候选实体的生成，如给定实体指称项“Michael Jordan”，实体链接系统根据规则和相关知识找到其可能指向的真实世界实体，如：“Michael B. Jordan”、“Michael Jordan (mycologist)”和“Michael Jordan (basketball player)”等。(2)候选实体的选择，系统根据实体的上下文及实体本身的知识，对所有的候选进行相似度的打分排序，根据排序的结果选择相应的候选实体作为链接对象。

为了验证本文提出的特征和方法的有效性，本文在 TAC KBP 2009 的实体链接评测数据上进行了测试。实验表明，本文提出的特征和方法在测试数据上显示出良好的性能。正确率达到 84.38%，高出参评队伍最好水平 2.21%。

本文章节安排具体如下：第二节介绍实体链接的相关工作；第三节介绍候选实体生成模块；第四节介绍候选实体选择模块；第五节为实验和结果分析；最后对本文工作进行了总结，并指出将来工作的方向。

2　相关工作

在命名实体消歧方面有很多关于实体链接的工作。Bagga [5] 等人用词袋模型来解决人名歧义的问题。Fleischman[6] 等人利用网络信息等特征训练最大熵模型来解决实体歧义问题。这些方法都是通过衡量指称项上下文文本与目标实体文本之间的相似度来判定两者是否一致。在这些方法中很大一部分都是利用词袋模型或者类似于词袋模型的方法，然而词袋模型只能捕捉表层字面匹配信息无法捕捉深层语义。

为了解决这一问题，Malin[7] 等人提出了利用随机游走的方法计算文本之间的相似度，除此之外，Han[8] 等人提出利用 Wikipedia 作为背景知识库，通过利用 Wikipedia 中的语义知识来进行消歧。利用不同的背景知识，研究者就可以得到不同的特征来进行实体消歧。

上面所述的方法中大多数都是解决单一实体链接问题，仅仅考虑目标实体与实体指称项间的语义相似度。除此之外 Cucerzan[9] 等为了更好地对于文本内的多个实体进行消歧，建立了全局语义约束，利用协同式策略综合考虑多个实体间的语义关联，从而进行协同实体链接。本文工作主要围绕解决单一实体链接问题开展。

如上所述，研究者们已经提出了很多不同的特

征用来进行实体消歧，如何有效合理地利用这些特征进行消歧也是一个研究热点，起初很多研究人员利用人工规则和权重来结合这些特征，然而这样不仅会耗费大量的时间和精力，还缺乏泛化能力。因此，有很多学者利用机器学习上的方法来完成特征的融合。Milne[10]等训练了很多类似SVM、C4.5和贝叶斯等典型的分类器来融合特征。这种基于分类的方法取得了不错的效果，但是该方法不能很好地考虑到候选实体之间的关系，为了解决这一问题本文利用排序学习的方法融合特征进而完成单实体链接的任务。

3 候选实体生成模块

为了完成实体链接，首先要从知识库中获得候选实体。在这一模块，我们为每个待消歧的实体指称项生成一组候选实体。通过对数据的分析不难发现，所有的候选实体应该在字面上和实体指称项相似，或者虽然字面上不相似但是实质上是同一实体的不同表示(如：别名或缩略名)。为了保证候选实体的高召回率，本文在候选生成阶段首先利用了基于表层字面信息的方法扩展指称项，将获取字面上最相近的一部分实体作为候选实体。接下来再利用Wikipedia针对每一个扩展后的词进一步扩展候选实体。具体方法如下：

3.1 基于表层字面信息的候选生成

在这一阶段主要是召回与实体指称项字面上相似的实体作为候选。首先我们利用Google开发的拼写错误修正工具来校正拼写错误，将可能的正确形式都作为候选实体加入候选列表中。接下来为了保证高召回率，本文又利用编辑距离计算实体指称项和知识库中每个实体间的相似度，经试验验证，本文选取编辑距离大于X的作为候选实体。按公式(1)计算编辑距离。

$$edit_{a,b}(i,j)=\begin{cases}\max(i,j) & \\ \min\begin{cases}edit_{a,b}(i-1,j)+1 & \\ edit_{a,b}(i,j-1)+1 & otherwise \\ edit_{a,b}(i-1,j-1)+1_{(a_i\neq b_j)} & \end{cases} & \end{cases} \tag{1}$$

除此之外，通过对数据的观察，我们发现在待消歧的实体指称项中有很多指称项是缩略词的形式，这种形式会造成很大歧义，但是其完全形式歧义很小，例如：

1) The ABC (Australian Broadcasting Corporation) is australia's national public broadcaster.

2) In American, ABC (American Broadcasting Company) first broadcast on television in 1948.

从例子中不难看出如果我们能从上下文中对指称项ABC进行扩展，得到其完全形式"Australian Broadcasting Corporation"和"American Broadcasting Company"，不仅能保证正确候选的召回，而且能减少候选实体与指称项之间的歧义，所以本文中采用了Zhang[11]等提出的缩略词扩展规则。

3.2 基于维基知识的候选生成

经过基于表层字面信息的候选生成，我们已经初步校正了拼写错误、得到了缩略词的全称和与实体指称项字面上相似的实体。但是还有很大一部分候选实体无法获得，因为有很多正确候选实体与实体指称项在字面形式上几乎完全不一致，如：实体"Michael Jordan"，如果在待消歧文本中出现了实体指称项"His Airness"，则通过基于表层字面信息的方法无法将实体"Michael Jordan"作为候选加入候选实体列表，因为表层字面匹配的方法缺乏深层语义知识，无法判断出"His Airness"是"Michael Jordan"的绰号。为了解决这一问题，在本文中我们引入了语义知识。本文挖掘并利用Wikipedia中的相关知识建立了实体指称项候选词典，用于补充基于表层字面信息生成候选的不足，以达到高召回率的目的。生成的部分实体指称项字典如表1所示。

表1 实体指称项字典实例

实体名	目标实体
IBM	IBM IBM mainframe IBM DB2 ...
AI	Artificial intelligence Game artificial intelligence Ai (singer) Angel Investigations Strong AI ...

本文利用Wikipedia中的以下信息建立词典：

• 重定向页面：自然界中很多实体的名字都不仅只有一个。这个问题也就是一义多词问题，在Wikipedia中用重定向页面来处理这类问题，同一

个实体只有一个实体页面，用这个实体流行度最广的名字作为标题。针对其余的名字都建立重定向页面，指向唯一的实体页面。所以重定向页面中包含很多的同义知识。

• 消歧页面：自然界中很多不同的实体具有相同的名字。这个问题也就是一词多义问题，在 Wikipedia 中用消歧页面来处理这类问题，消歧页面中有一系列的链接信息，分别链向这个名字所指的不同实体。所以能从消歧页面中发现很多候选实体。

• 锚文本信息：在 Wikipedia 的文本中会将提到的重要实体链接到相对应实体的页面，这就是锚文本信息，这些锚文本有的是对应实体的同义词，有的是对应实体的别名，还有的是对应实体的名字的错误拼写。能为候选的生成提供重要依据。

为了测试本文提出的候选生成模块的召回率，我们在 TAC KBP 2009-2013 年的数据上都进行了测试。测试结果如表 2 所示。测试数据表明本文提出的候选生成模块的有效性。

表 2　候选实体生成模块召回率测试

数据集	TAC2009	TAC2010	TAC2011	TAC2012	TAC2013
召回率	0.947 7	0.925 3	0.933 3	0.918 9	0.867 5

4　候选实体选择模块

基于候选实体生成模块，我们可以得到一个实体指称项对应的所有的候选实体，为了正确链接实体，我们必须对所有的候选实体排序，最终将得分最高的实体作为此实体指称项的链接实体。本文对候选实体的选择基于一个有监督的排序学习算法。对于一个实体指称项，排序学习分类器的输入是 n 个 d 维空间向量，其中 n 表示的是该实体指称项的候选实体的数目，每一对候选实体与实体指称项会根据特征函数生成一个 d 维空间的向量，其中 d 代表特征的个数，这些特征充分考虑了候选实体自身的信息以及指称项上下文内容与候选实体的语义相似度等知识。通过最大边缘化的方法来选择候选实体，即正确的实体所获得的分数应该高于其他的候选实体的分数同时加上一定的余量。这个约束条件等同于 SVM 排序学习算法[12]，优化函数和约束条件为式(2)～(4)。

$$\text{Minimize: } V(w,\varepsilon) = \frac{1}{2}w*w + C\varepsilon_{q,k} \tag{2}$$

$$\text{Subject to: } W(\phi(q,q.e) - \phi(q.e_k)) \geqslant 1 - \varepsilon_{q,k} \tag{3}$$

$$\varepsilon_{q,k} \geqslant 0 \tag{4}$$

其中，V 为损失函数，w 是要学习到的关于特征的权重，c 为惩罚因子。$q.e$ 是实体指称项的正确实体，$q.e_k$ 实体指称项的其他候选实体。约束条件的物理意义是正确实体获得的分数要尽量大于其他候选实体获得的分数。本文的候选实体选择模块共使用了表层字面特征、深层语义特征和空实体特征共三类七种。各个特征将在下面进行详细介绍。

4.1　表层字面特征

这类特征主要从表层字面信息考虑候选实体与待消歧的实体指称项间的相似度。这类特征包括编辑距离相似度、Dice 相似度、向量空间相似度和实体共现信息等特征。为了提高特征的有效性，在计算表层字面特征时我们会对文本进行预处理。具体如下：

数据预处理：在数据预处理阶段，本文会过滤掉实体名字中的括号及括号中的内容，还会考虑到缩略词和大小写的情况。

基于编辑距离的相似度 Edit：该特征主要用来度量候选实体名和待消歧指称项的编辑距离，编辑距离的计算如上述公式(1)所示。

基于 Dice 系数的相似度 Dice：该特征主要用来衡量候选实体名和待消歧指称项的 Dice 系数。如：x 和 y 为两个字符串，则 Dice 的计算如公式(5)所示。

$$S = \frac{2n_t}{n_{x+}n_y} \tag{5}$$

公式中 n_t 表示同时出现在字符串 x 和 y 中的二元组个数，n_x 是字符串 x 中的二元组个数，n_y 是字符串 y 中的二元组个数。

基于向量空间模型的篇章级相似度 Bow：该特征主要用来衡量待消歧指称项的上下文文本和候选实体的描述文本之间的相似度。同一实体出现的上下文环境应该类似，所以这一特征在传统的消歧方法中占有很重要的地位。计算时，应先将待消歧指称项的上下文和候选实体的上下文用词袋子模型表示成向量，向量中的每一维都由标准的 TF-IDF 计算得到，最后计算向量的余弦值作为相似度。

实体共现信息 Co：该特征为二元特征，标志着指称项实体名是否在候选实体文本中出现，或者候选实体名是否在指称项上下文中出现，出现则设为 1，否则为 0。

4.2 深层语义特征

在上述的表层字面信息特征中，主要是基于词或者实体的匹配信息，无法捕捉到深层语义，对于上下文中匹配信息较少的情况不具备泛化能力，所以我们应当将深层语义信息考虑进来，在深层语义特征中，主要是利用从 Wikipedia 中获得的背景知识计算候选实体与实体指称项之间的深层语义关联。具体如下：

实体流行度 Pou：这个特征主要是衡量一个实体在一篇文章出现概率的大小。本文中我们的计算方法同 Han[13] 一样，统计出实体在整个知识库中出现的总次数 N，再统计被链接到的次数 L，则 L/N 为实体的流行度。例如：给出实体“Michael Jordan”，在没有其他任务附件信息的条件下，从流行度可以知道，实体“Michael Jordan”链向美国著名篮球明星“Michael Jeffrey Jordan”的概率要大于链向伯克利大学教授“Michael I. Jordan”的概率。

基于维基实体的相似度 Ws：为了更准确地计算实体指称项和候选实体之间的相似度，本特征使用 Wikipedia 知识来获取实体之间的语义关系。类似于 Han[8] 等的工作，本文中实体相似度计算分为三步：①指称项文本中的 Wikipedia 实体向量表示的抽取；②两个指称项实体向量表示的对齐；③相似度计算。以下分别具体介绍：

① 指称项文本中的 Wikipedia 实体向量表示的抽取：为了计算实体之间的相似度，首先将每个实体 e 表示成 Wikipedia 的实体向量 $e=\{(c_1,w(c_1,\mathrm{e})),(c_2,w(c_2,\mathrm{e})),\cdots,(c_m,w(c_m,\mathrm{e}))\}$，其中，$C_i$ 是指称项上下文中的 Wikipedia 实体，而 $w(C_i,\mathrm{e})$ 是实体 C_i 在指称项 e 的实体向量表示中的权重。给定实体，其实体向量表示的抽取分两步完成：首先完成 Wikipedia 的实体抽取，本文利用由 Milne[14] 等开发的工具 Wikipedia-Miner 来识别并抽取，同一个指称项文本中抽取的实体集合组成向量。还要为向量中的每维估计权重，因为不同的实体在消歧过程中起的作用是不一样的，本文中，一个实体 c 在一个实体指称项 e 中的重要性，计算如公式(6)所示。

$$w(c,\mathrm{e})=|e|-1\left(\sum_{c_i\in e,c_i\neq c}sr(c,c_i)\right) \tag{6}$$

其中 $sr(c,c_i)$ 是 Milne[10] 提出的 Wikipedia 实体之间的语义关联。根据实体权重，我们可以过滤掉噪音实体从而提升实体消歧系统的效率和性能。

② 实体向量表示的对齐：将指称项用 Wikipedia 实体向量表示后，我们可以使用余弦相似度等传统方法来计算实体之间的相似度。但是传统相似度通常不能考虑到实体之间的语义关联。因此，我们利用实体对齐方法来识别实体之间的对应关系，并以此为基础来在实体相似度计算中融入实体之间的语义关系。给定两个实体的向量表示 e_l 和 e_k，我们使用如下方法实现向量中实体的对齐：对 e_l 中的每一个实体 c，我们选择目标实体 e_k 向量表示中与其有最大语义关联度的实体作为它的对齐实体，计算如公式(7)所示。

$$Align(c,e_k)=\underset{c_i\in e_k}{\operatorname{argmax}}sr(c,c_i) \tag{7}$$

③ 相似度计算：完成实体对齐后，指称项相似度计算的关键问题是如何将这些实体对齐信息结合到相似度计算中，进而在相似度计算中融入 Wikipedia 语义知识。基于实体对齐的结果，我们认为从一个实体 e_l 到另一个实体 e_k 的语义关联为“两个实体之间所有对齐实体之间语义关联的带权平均”，计算如公式(8)所示。

$$SR(e_k\rightarrow e_l)=\frac{\sum_{c\in e_k}w(c,e_k)\times w(Align(c,e_l),e_l)\times sr(c,Align(c,e_l))}{\sum_{c\in e_k}w(c,e_k)\times w(Align(c,e_l),e_l)} \tag{8}$$

按上述定义，给定两个实体 e_l 和 e_k，从 e_l 到 e_k 的和 e_k 到 e_l 的语义关联度是非对称的。因此本文利用的两个实体 e_l 与 e_k 之间的相似度为 e_l 到 e_k 和 e_k 到 e_l 的语义关联度的平均值。

经过上述的三步，我们可以计算出两个实体的深层语义相似度，更好地捕捉实体之间的深层语义信息。

4.3 空实体特征

除了上述的两类特征，为了更好地处理空实体的问题，本文参照 Dredze[15] 等人的工作，设计了空实体特征 NIL，并且在候选特征中强制加入空实体作为候选，一同参与所有候选实体的打分排序，如果是空实体得分最高，则将待消歧的实体指称项链接到空实体上。在本文中空实体的特征向量中只有空实体特征设定为非 0，其余的特征值均为 0。对于其他的候选实体来说空实体特征为 0。其余的特征由上述的定义计算得到。

5　实验结果及分析

5.1　实验数据集

本文的实验在 TAC KBP 2009 的评测数据集上进行。TAC KBP 评测中实体链接的任务目标是将文本中的实体指称项与目标实体知识库中的相应实体链接。TAC KBP 2009 实体链接任务由目标实体知识库和评测数据两部分组成评测数据集，如下所示：

目标实体知识库：评测任务中，指称项目标实体的相关信息存储在目标实体知识库中。目标实体知识库以实体为单位组织。目前，TAC 实体链接任务知识库中的目标实体是从 2008 年 10 月的 Wikipedia 中抽取构建，在目标实体知识库中，每一个实体节点包含如下几方面信息：知识库 ID、实体的类别、实体的名字、属性信息和消歧文本。

评测数据：评测任务中，测试数据 Query 以 XML 格式提供，每个 Query 能提供的信息有实体指称项的名字、Query ID、实体所在文本的 ID 和实体指称项在实体知识库中的对应实体的 ID。评测数据中总共包括了 3 904 个 query，其中 2 229 个 query 是在知识库中找不到对应的实体的，也就是要标记为空实体，其余的 1 675 个 query 能在知识库中找到指定的实体。在评测数据中不同类别的实体，其中 627 个关于 PER 的 query，2 710 个关于 ORG 的 query，567 个关于 GPE 的 query。

5.2　评价指标

本文采用 TAC KBP 2009 中的评测指标 Micro-averaged accuracy 来评价实体链接的效果，计算如公式(9)所示。

$$Micro = \frac{\sum_{q \in Q} \sigma(L(q), C(q))}{|Q|} \tag{9}$$

公式(9)衡量了所有链接结果的平均准确率，其中 $L(q)$ 是实体链接系统给出的 query q 的目标实体 ID，Q 是所有 query 的集合，$C(q)$ 是 query q 的准确目标实体 ID，$\sigma(L(q), C(q))$ 用于判断 $L(q)$ 是否与 $C(q)$ 相同，不相同为 0，相同则为 1。

5.3　实验设置

在实验中我们首先不考虑空实体，仅在知识库中能找到相应实体的数据集上进行实验，之后在候选实体中加入空实体，在特征中加入空实体特征，在整个数据集上进行实验，验证考虑空实体后本文提出的实体链接系统的性能。与 Shen[16] 等一样，本文的所有实验数据都是在 TAC KBP 2009 的数据集上采用十折交叉验证获得的。

5.4　结果及分析

5.4.1　特征有效性分析

为了验证本文利用的深层语义知识特征的有效性，我们将第 4 节提出的特征进行了不同方式的组合进行实验。为了降低空实体对实验的影响，实验时我们只在能在知识库中找到实体的 1 675 个 query 的数据集上进行测试。实验结果如表 3 所示。

表 3　在非空实体上的测试结果

编号	实验特征	Micro-averaged accuracy
#1	Bow	0.780 8
#2	Ws	0.791 1
#3	Edit+Dice+Bow+Co	0.841 7
#4	Ws+Pou	0.854 3
#5	Edit+Dice+Bow+Co+Ws+Pou	0.894 9

从实验结果上看，#2 与 #1 对比性能提升 1% 左右，说明在实体链接的过程中单独运用基于维基实体的相似度效果会优于单独利用向量空间模型的相似度。证实了基于维基实体的相似度能更好的捕捉实体之间的语义信息。

#3 与 #1 相比性能提升 6% 左右，相较于 #1 系统，#3 系统又融入了三个基于表层字面信息的特征，分别是基于编辑距离的相似度、基于 Dice 系数的相似度和实体共现信息。结果性能上的提升说明了本文提出的这三个特征的有效性，也说明了在实体消歧阶段，实体名相似与否或者实体名是否共现具有很重要的地位。

#4 与 #2 的比较中可以看出特征实体流行度的有效性。流行度表征了一个实体在一篇文章中出现的概率，这说明一个实体的流行度越大，那么当它作为候选实体进行消歧时，其被判定为正确答案的概率也就越大。

#4 的性能要优于 #3，#4 中的特征都是基于 Wikipedia 中的超链接等关系获得的深层语义知识，而 #3 中的特征都是基于字面表层信息的。实验结果说明基于 Wikipedia 获得的知识能捕捉更多

的语义知识。更有利于实体消歧的效果提升。

#5的效果明显优于其余的对比实验，这说明基于Wikipedia获得的语义知识与基于表层字面特征能捕获的知识是互补的，单独的应用二者都不能达到最优效果，应当将两类特征结合应用。

5.4.2 算法有效性分析

为了验证本文利用的基于排序学习算法框架的候选实体选择方法优于传统的分类方法，本文实现了基于Naive Bayes的分类方法和基于SVM的分类方法来进行对比实验，最终实验分别在不同的特征组合下进行，用分类的方法时，首先将所有的候选实体分类为两类，取正确的类别中概率最高的为最终链接的对象，具体结果如图1所示。

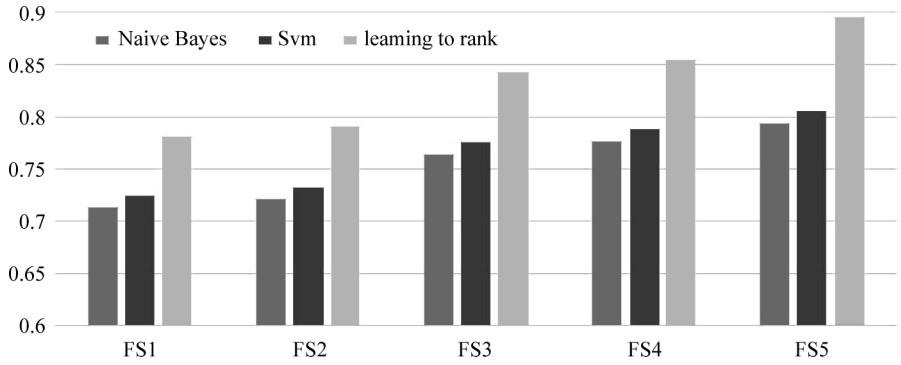


图1 基于贝叶斯分类器、Svm分类器和基于排序学习的实体链接系统效果对比图

如图，实验在五个特征集合上进行，此实验选择的五种特征组合与测试特征有效性实验中的五个特征组合一致，如：FS1(feature set1)中的特征与#1中的特征一致。在五个集合上利用排序学习算法选择候选实体的效果都要明显优于利用传统分类器的效果。在FS1上性能提升最少，从0.713 4提升到0.780 9，提高6%左右，在FS5上性能提升最高，从0.794提升到0.894 9，提升10%左右。实验结果表明，基于排序学习框架的算法更适合实体链接任务。

5.4.3 与state-of-Art系统性能对比

上述实验都是针对在知识库中能找到实体的情况进行的。为了测试系统在完整数据集上的性能，我们在候选实体中加入空实体，在特征中加入空实体特征。在TAC KBP 2009的完整数据集上进行十折交叉测试。结果与参加TAC KBP 2009的前三名[17]进行比较，结果如表4所示。

表4 系统整体性能测试和TAC KBP 2009的前三名系统对比

系统	accuracy of all queries	accuracy of non-NIL queries	accuracy of NIL queries
Siel 09	0.821 7	0.765 4	0.864 1
QUANTA	0.803 3	0.772 5	0.824 1
hltcoe	0.798 4	0.706 3	0.894 1
our	0.843 8	0.798 2	0.877 8

从上述实验结果可以看出，本文构建的系统在性能上达到84.38%，高出参加评测的最好成绩2.21%。不仅说明了本文构建的实体链接系统的可靠性，也说明了本文利用的特征和方法的有效性。

6 总结与展望

本文针对传统实体间相似度计算方法存在的不足，利用了一种基于深层语义知识计算实体之间相似度的方法。为了更好地融合多种特征，本文设计了一个基于排序学习算法框架的实体链接系统。实验结果表明，相比于传统的计算方法，新的相似度计算方法可以更加有效地捕捉实体指称项文本与候选实体间的语义关联。同时，融入了多种特征的实体链接系统的性能在TAC KBP 2009的数据集上取得了良好的性能，正确率达到84.38%，高出参加评测的最好成绩2.21%。

下一步的工作主要包括：1)本文建立的实体链接系统对空实体的处理还不完善，仅仅是指出该实体指称项所表示的实体在知识库中不存在，还需要将这项工作进行细化，如将空实体进行聚类并且将聚类后的空实体加入到知识库中；2)尝试使用其他的排序学习算法，如Listnet[18]等。

参考文献

[1] 赵军，刘康，周光有等. 开放式文本信息抽取[J]. 中文

信息学报，2011，25(6)：98-110.

[2] Fabian M Suchanek，Gjergji Kasneci，Gerhard Weikum. Yago：A large ontology from wikipedia and wordnet. Web Semantics：Science，Services and Agents on the World Wide Web，2008，6(3)：203-217.

[3] Fei Wu，Daniel S Weld. Automatically refining the wikipedia infobox ontology[C]//Proceedings of the 17th international conference on World Wide Web，2008：635-644.

[4] Sören Auer，Christian Bizer，Georgi Kobilarov，et al. Dbpedia：A nucleus for a web of open data. In The semantic web，2007：722-735.

[5] Amit Bagga，Breck Baldwin. Entity-based cross-document coreferencing using the vector space model[C]//Proceedings of HLT/ACL，1998：79-85.

[6] Michael B Fleischman，Eduard Hovy. x Multi-document person name resolution[C]//Proceedings of ACL，Reference Resolution Work shop，1998：66-82.

[7] Bradley Malin，Edoardo Airoldi，Kathleen，et al. A network analysis model for disambiguation of names in lists[J]. Computational & Mathematical Organization Theory，2005，11(2)：119-139.

[8] Han X，Zhao J. Named entity disambiguation by leveraging Wikipedia semantic knowledge[C]//Proceeding of the 18th ACM conference on Information and knowledge management，2009：215-224.

[9] Silviu Cucerzan. Large-scale named entity disambiguation based on wikipedia data[C]//Proceedings of EMNLP CoNLL，2007，7：708-716.

[10] David Milne，Ian H Witten. Learning to link with wikipedia[C]//Proceedings of the 17th ACM conference on Information and Knowledge Management，2008：509-518.

[11] Tao Zhang，Kang Liu，Jun Zhao. The nlprir entity linking system at tac 2012.

[12] Thorsten Joachims. Optimizing search engines using clickthrough data[C]//Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2002：133-142.

[13] Xianpei Han，Le Sun. A generative entity mention model for linking entities with knowledge base[C]//Proceedings HLT/ACL，2011：945-954.

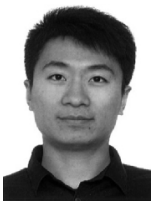
[14] David Milne，Ian H Witten. An open-source toolkit for mining wikipedia[J]. Artificial Intelligence，2013，194：222-239.

[15] Mark Dredze，Paul McNamee，Delip Rao，et al. Entity disambiguation for knowledge base population[C]//Proceedings of CL. 2010：277-285.

[16] Wei Shen，Jianyong Wang，Ping Luo，et al. Linden：linking named entities with knowledge base via semantic knowledge[C]//Proceedings of WWW，2012：449-458.

[17] Paul McNamee，Hoa Trang Dang. Overview of the tac 2009 knowledge base population track. In Text Analysis Conference (TAC)，2009，17：111-113.

[18] Zhe Cao，Tao Qin，Tie-Yan Liu，et al. Learning to rank：from pairwise approach to listwise approach[C]//Proceedings of ICML，2007：129-136.



陈玉博(1990—)，**博士，主要研究领域为事件抽取、信息抽取和自然语言处理。**
E-mail：yubo.chen@nlpr.ia.ac.cn

刘康(1981—)，**博士，副研究员，主要研究领域为信息抽取、网络挖掘、问答系统等。**
E-mail：kliu@nlpr.ia.ac.cn

何世柱(1987—)，**博士，助理研究员，主要研究领域为智能问答、知识工程以及自然语言处理。**
E-mail：shizhu.he@nlpr.ia.ac.cn