文章编号：1003-0077（2011）00-0000-00

基于多源知识和 Ranking SVM 的中文微博命名实体链接

陈万礼，眷红英，吴泳钢

（郑州大学 信息工程学院，河南 郑州 450001）

摘要：命名实体在文本中是承载信息的重要单元，正确分析存在歧义的命名实体，对文本的理解起到关键性作用。本文提出基于多源知识和 Ranking SVM 的中文微博命名实体链接，结合同义词词典、百科资源等知识产生初始候选实体集合，同时从文本中抽取多种组合特征，利用 Ranking SVM 对候选实体集合进行排序，从而得到目标实体。在 NLP&CC2014[1]中文微博实体链接评测数据集上进行了实验，获得了 89.40%的微平均准确率，与 NLP&CC2014 中文微博实体链接评测取得最好成绩的系统相比，本文的系统具有一定的优势。

关键词：命名实体；中文微博实体链接；同义词词典；百科资源；Ranking SVM；语义特征

中图分类号：TP391　　**文献标识码：**A

Chinese Micro-blog Named Entity Linking Based on Multisource Knowledge and Ranking SVM algorithm

CHEN Wanli, ZAN Hongying,WU Yonggang

(School of Information Engineering, Zhengzhou University, Zhengzhou Henan, 450001, China)

Abstract: Named entity is an important component conveying information in texts; so an accurate understanding of the named entities is needed to ensure a correct analysis of the text information. This paper proposes a Chinese micro-blog entity linking strategy based on multi-resource knowledge and Ranking SVM algorithm, combing a dictionary of synonyms, the encyclopedia resources to produce an initial set of candidate entities , then various combinations of features extracted from the text and the use of Ranking SVM algorithm to generate the second sort of candidate entity set. In this strategy, named entities to be linked in micro-blog are mapped to the corresponding candidate entities in the knowledge base. The evaluation results gain a micro average accuracy of 89.04%, based on experiments using data sets of NLP&CC2014 Chinese micro-blog entity linking track. Compared with the state-of-the-art result, the accuracy of this method demonstrates the effectiveness of our method.

Keywords: named entity; chinese micro-blog entity linking; dictionary of synonyms; encyclopedia resources; Ranking SVM; semantic features

1 引言

据《第 35 次中国互联网络发展状况统计报告》[1]显示，截至 2014 年 12 月，中国网民规模达 6.49 亿，其中手机网民规模 5.57 亿，互联网普及率达到 47.9%。由此可见互联网规模之大，已经成为人们生活的重要部分。而这种爆炸式的增长带来的问题之一便是用户产生的内容数据急剧增长，其中大多数为文本数据，进而促使了文本方面的大数据分析技术的广泛使用。而这种分析挖掘必然面临对于词义正确理解的强烈需求。由此可见，解决命名实体链接问题的必要非常之大，将存在歧义的实体正确地链接到对应的知识库中，具有重要意义。

本文实验数据来自新浪微博，而微博与普通文本最显著的区别在于内容长度限制在 140

[1] http://tcci.ccf.org.cn/conference/2014/index.html

* **收稿日期：　　定稿日期：**

基金项目：国家自然科学基金项目（61402419，60970083，61272221）；国家社会科学基金项目（14BYY096）；国家高技术研究发展 863 计划（2012AA011101）；河南省科技厅科技攻关计划项目（132102210407）；河南省科技厅基础研究项目（142300410231，142300410308）；河南省教育厅科学技术研究重点项目（12B520055，13B520381）；计算语言学教育部重点实验室（北京大学）开放课题（201401）；国家重点基础研究发展计划 973 课题（2014CB340504）；河南省高等学校重点科研项目（15A520098）。

字以内,发布的内容具有如下特点:文本长度短、口语化、表达不清晰等问题.因此，对应于上述特点，针对微博数据的命名实体链接也面临一些新的问题。

针对微博上述特点,本文提出了基于多源知识和Ranking SVM的中文微博命名实体链接，主要包括以下几个方面：1）对知识库进行更新，添加实体对应的中文维基百科分类，并且分别从中文维基百科、互动百科、百度百科抽取实体的别称（同义词），以此来提高实体链接的准确性；2）采用百度搜索引擎对存在错别字的待链接命名实体（简称：目标实体）名称进行纠正；3）利用Lucene[2]对知识库中的所有候选实体建立本地索引，根据微博中的命名实体检索得到初始候选实体集合；4)抽取候选实体集的语义特征，利用训练得到的Ranking SVM模型对初始候选实体排序；5）从候选实体中找出得分最高的实体，如果符合相关条件，则返回KB_ID；否则，返回NIL。

2 相关研究

命名实体链接的输入为一段文本，称为查询文档。查询文档包含诸多实体名称，称为查询名称。而命名实体链接的目的则是从指定知识库中找到查询名称所指代的实体[2]。

命名实体链接任务通常包括两个主要阶段：候选实体生成与候选实体歧义消解。候选实体生成主要是对查询词语的扩展，另外对待链接实体的上下文特征的抽取，也属于候选实体集合的初步生成环节；而候选实体的歧义消解则是对初步生成的集合进行排序，以确定最优选项。命名实体链接的任务可以归纳为图1所示流程。

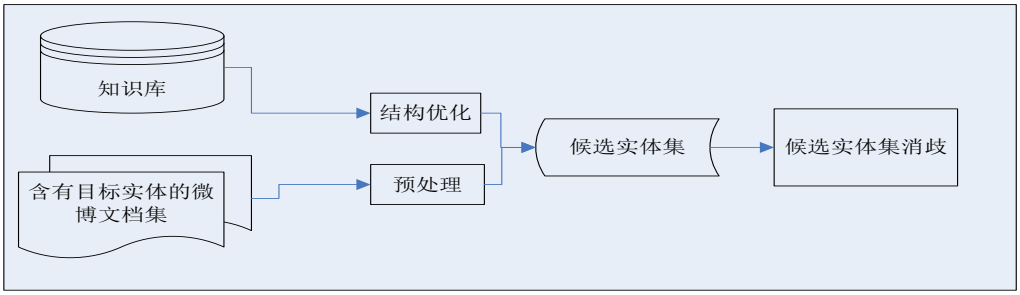

图1 命名实体链接基本流程图

2.1 候选实体生成方法

候选实体生成主要是得到知识库中和查询名称相关联的初始候选实体集,其目的是为了缩小知识库的实体歧义消解范围。在获取初始候选实体集时， Mihalcea和Csomai[3]，Milne和Witten[4]从Wikipedia中抽取以查询名称为锚文本的文本片段,并进一步找到超链接目标页面对应的实体作为候选。Bunescu 和 Pasca[5]，Cucerzan[6]则为候选实体生成专门构造了命名实体词典。Gottipati等采取对查询名称的扩展[7]，Sun利用Wikipedia的重定向页面、消歧页面等建立词典[8]，为解决查询名称为缩略语的实体生成问题，Zhang[9]提出了在查询文档中查找对应全称词语的方法。总之，前人通过上述多种方法，在一定程度上提高了系统召回率。

2.2 候选实体消歧方法

候选实体歧义消解则主要是从初始候选实体集中选出最有可能的目标实体,进而将查询名称链接到目标实体。Varma[10]等人利用搜索引擎工具对候选实体进行排序，以此选出相似度分值最高的候选实体，Han和Zhao[11]通过BOW（bag-of-words）与Wikipedia的语义网络对候选实体进行相似度计算，Zheng[12]等人提出了 L2R（Learning to rank）算法进行实体消歧，Zhang[13]等人利用 SVM（Support Vector Machine）对候选实体进行分类，以达到消歧目的。

本文参考上述方法及微博特点，提出了同义词词典、百科知识和Ranking SVM模型相结

[2] http://lucene.apache.org/

合的策略来获取知识库候选实体。

3 命名实体链接

本文把命名实体链接任务分成两个阶段，即候选实体生成和候选实体歧义消解，针对候选实体生成主要采用构建多源知识的方法来完成，而候选实体歧义消解部分则采用有监督的 Ranking SVM 模型来对候选实体进行排序消解。

3.1 候选实体生成

3.1.1 数据预处理

本文实验数据来自新浪微博，而知识库则是维基百科，由于微博内容的长度被限定在 140 个字符之内，并且发布内容具有文本长度短、口语化、表达不清晰等特点。相应地，对于新浪微博的研究主要有如下问题：

a) 微博的内容构成复杂，常常出现“#”符号，两个“#”符号之间的内容为话题，还有汉字形式的表情（比如：[衰]、[高兴]等）；以及“@”符号等。

b) 外来音译名称，如“萨科齐”与“萨柯奇”等，本质上二者指代的为同一人物即法国前总统“Nicolas Sarkozy”等。

c) 微博内容的表达口语化，经常出现错别字。

d) 微博内容中繁体汉字、简体汉字、拼音的结合出现，如：“鄭州”、“fudan 大學”等。

通常，我们需要对诸如上述问题中的部分情况，进行预处理，比如 a 中的微博符号，可以制定相应的规则对微博文本中的符号进行处理，而对于 b 和 c 中的外文名称音译问题，则利用百度搜索引擎提供的“候选词推荐”功能来辅助降低问题复杂性，对于 d 中提到的情况，则采用繁简字体转换和拼音汉字转换的方法进行处理，最终统一为汉字简体形式。

3.1.2 同义词表构建

微博内容中目标实体的表达形式具有多样性，包括别名、简称及绰号等，根据 Han[14] 等人的统计，在 Tweets 中每个命名实体平均有 3.3 个不同的表达形式，为了处理表达形式多样性的问题，本文从维基百科（中文版）、互动百科、百度百科获取候选实体对应的所有实体信息，进而将实体对应的不同表达形式进行归纳总结，构建同义词表，以此提高命名实体链接的准确性，从维基百科、互动百科及百度百科分别对知识库中的 8405 个、5492 个及 6235 个实体进行了同义词扩展。以“沙奎尔·奥尼尔”为例（如表 1 所示），同义词的扩展采用模式匹配方式，比如以百科源代码网页中的“绰号”、“nickname”、“别名”标签作为基本匹配模版进行同义词的抽取，详见 2012 年 CCF 自然语言处理与中文计算会议中关于中文词汇语义关系抽取[3],通过构建同义词表，我们对于图 1 中的待链接实体“大鲨鱼”和“大柴油机”时，可以准确而快速地找出其中文实体全称为“沙奎尔·奥尼尔”（英文实体全称为“Shaquille O'Neal”）。

表 1 “沙奎尔·奥尼尔”对应的维基百科、互动百科的同义词（别名）

知识源	同义词
维基百科	Shaq、The Diesel、大鲨鱼、侠客、扎克博士、大仙人掌
互动百科	大鲨鱼、奥胖、超人、大柴油机
百度百科	大鲨鱼、超人

图 1 待链接实体为别名的微博示例

```
- <weibo id="aonierqiuyituiyi944">
    <content>【奥尼尔球衣高悬斯台普斯 比肩传奇一生湖人】http://t.cn/zT2BOgB 为湖人效力八年、拿下三连冠的大鲨鱼、超人、大柴油机、亚里士
      多德、奥尼尔的球衣在今天退役了，斯台普斯中心，科比为奥尼尔送上祝福，珍妮巴斯和禅师菲尔杰克逊也在现场见证这一时刻！</content>
    <name id="1">大鲨鱼</name>
    <startoffset id="1">56</startoffset>
    <endoffset id="1">59</endoffset>
    <kb id="1" />
    <name id="3">大柴油机</name>
    <startoffset id="3">63</startoffset>
    <endoffset id="3">67</endoffset>
    <kb id="3" />
  </weibo>
```

[3] http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html

3.1.3 人物 Title 表的构建

人物实体的 title 主要是社会地位，自身社会关系以及从事职业的一种标识，比如：“发言人”、“歌手”、“公安局长”等，而这些 title 在对人物进行实体消歧时，可以辅助区分具有相同名称的不同实体。论文所采用的 title 词汇资源为 HowNet 中一部分，共计 244 个，如表 2 所示。例如，名称“李娜”，在百度百科中“李娜”对应 26 个义项，其中 25 个义项为人物实体名称（如图 3.2 所示），可以通过在“李娜”实体所在文档中的 title 描述来为实体消歧提供有效的特征信息。

表 2 部分人物 title 列表

职称列表（部分）	澳督 班长 编播员 编辑 编剧 辩护律师 部长 裁判 采煤工 参谋长 藏学家 常委……

表 3 百度百科“李娜”的部分义项

李娜	中国女子**网球名将**
	南开大学医学院**副教授**
	潮州市**政协副主席**
	……

3.2 候选实体歧义消解

根据初始候选实体抽取指定的特征组，利用训练得到的 Ranking SVM 模型对待链接实体和候选实体集合构造的特征文本进行预测，具体过程见算法 1。

算法 1 基于 Ranking SVM 模型的实体链接

输入：目标实体(mention)，微博内容，知识库

输出：知识库中候选实体的 KB_ID，或者 NIL

1. Begin
2. 选择 mention 所在微博的最小子句（以句号，逗号，感叹号为结束）;
3. 通过 NLPIR[4]工具，将微博句子进行分词、词性标注，经过停用词过滤，抽取特征;
4. similarity:=-1, index:=-1;
5. 将步骤 3 得到的待检索词组在已经建立索引的知识库中进行检索，得到检索结果集合 condidate _entities;
6. For e_i ∈condidate_entities
7. Begin
8. 抽取 e_i 所在的知识库文本和待链接实体的特征，构造特征组合 $feature_i$;
9. End
10. 利用训练得到的 Ranking SVM 模型对 condidate_entities 构造的特征文本进行预测，得到对各个候选实体的分值 γ_i ， score[i] := γ_i;
11. For γ_i∈score
12. Begin
13. 如果 γ_i > silimarity,则 index :=i ;
14. End
15. 如果 similarity>λ（λ 为预先设置的阈值），则输出 KB_ID;
16. 否则 输出 NIL
17. End

[4] http://ictclas.nlpir.org/

3.2.1 Ranking SVM 模型

Ranking SVM 模型是由 Herbrich[15]等人提出的一种排序算法，它可以广泛地应用于信息检索领域，如 Cao[16]等人利用此类模型进行文档检索任务。Joachims[17],[18]等人提出了基于 Paiwise 的数据标注方法，并提供了免费的 SVMrank 工具[5]。

假设存在一组输入向量 $X \in R^n$，n 在此表示特征的维数，同时存在一组输出向量 $Y = \{r_1, r_2, r_3, \mathrm{L}\ \mathrm{L}\ , r_n\}$，n 表示排序数。进一步假设存在一组全序排列 $r_n > r_{n-1} > \mathrm{L}\ \mathrm{L}\ > r_1$，“>”符号表示一种优先权的偏向关系，那么将存在一系列排序函数 $f \in F$ 决定了下列的偏序关系：

$$S' = \left\{ x_i^{(1)} - x_i^{(2)}, z_i \right\}_{i=1}^{l} \tag{1}$$

Herbrich 等人将上述的排序学习问题看做基于实例对的分类学习问题。首先设定一个线性函数 f 。

$$f(x; w) = \langle w, x \rangle \tag{2}$$

其中，w 表示一组权重向量，“,”表示向量的内积。

通过公式 1、2 可以得到，如下关系：

$$\langle w, x^{(1)} - x^{(2)} \rangle > 0 \Leftrightarrow f(x^{(1)}; w) > f(x^{(2)}; w) \tag{3}$$

将公式 3 转换为二值分类问题，则可以表示为：

$$(x^{(1)} - x^{(2)}, z), z = \begin{cases} +1 & x^{(1)} \mathrm{f}\ x^{(2)} \\ -1 & x^{(2)} \mathrm{f}\ x^{(1)} \end{cases} \tag{4}$$

对于给定的训练数据 S，我们以此构造一个新的包含 l 个向量的训练数据集合 S'，将 S' 中的数据作为分类数据构造 SVM 模型，对任意一组向量 $X^{(1)} - X^{(2)}$ 赋以分类类别，其中 z =+1 代表正样例，z =-1 代表负样例。后续问题即转化为二次最优化问题，具体可参考[19]。

3.2.2 Ranking SVM 模型的特征选择

本文选定了三类特征，分别是表面性特征（实体流行度、是否子串、是否满足编辑距离阈值）、上下文的文本相似度特征、主题相关性特征。

(一) 表面性特征

[1].实体流行度

提取实体流行度特征即求 query 对应的候选实体中概率最高的实体。这个概率可以通过很多方式计算得到，比如：计算 query 作为超链接指向各个候选实体的链接比例来获得。因此，如果一个 query 对应的候选实体集为 $E = \{(e_1, C_1), (e_2, C_2), (e_3, C_3), \mathrm{L}\ , (e_n, C_n)\}$，其中 C_i 是实体 e_i 对应的在超链接中被指向的次数。则该候选实体的流行度 P 为：

$$P = \frac{C_i}{\sum_{k=1}^{n} C_k} \tag{5}$$

[2].候选实体与待链接实体之间是否属于子串的关系

[5] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

子串关系特征在此主要指一个字符串是另外一个字符串的开头或者结尾，而对于如人名“Michael Jeffrey Jordan”、“Michael Jordan”这样的子串关系，识别难度还是比较大的。这种情况下，可以使用下述 *Dice* 系数方法来识别。本文中子串关系特征主要是针对子串是母串的开头或者结尾的情况。

[3].候选实体与待链接实体之间编辑距离是否小于设定阈值

英文中人名通常不写中间那个名字，比如“Michael Jeffrey Jordan”常常写为“Michael Jordan”，还有英文中大量组合词，如“home-made”也会被写成“homemade”，其实它们所指的是一样的。在此，计算方法可以使用编辑距离或者 Dice 系数等，而阈值设定则需要通过实验来调优。

（二）上下文文本相似性特征

本文的内容相似性衡量是先将上下文文本转换成文本向量，利用向量空间模型计算文本向量相似性。空间向量的相似度有如下计算方法：

[1].余弦相似度

余弦相似度是通过计算两个向量在空间中的夹角余弦值来衡量彼此之间的相似程度，取值范围在[-1,+1]，余弦相似度是计算相似度的常见方法，类似的还有 Dice 系数，Jaccord 系数，如果有向量 A 和 B，其向量之间夹角记为θ，则其计算如公式 6 所示。

$$D=\cos(\theta)=\frac{A\cdot B}{|A||B|}=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{6}$$

[2].欧几里得距离的相似性

相似度的衡量除了可以计算相似性，还可以计算它们的不相似性，比如计算它们之间的距离，距离大，相似度就小。在距离衡量常用的是欧几里得距离，也叫欧氏距离，它主要是计算空间中两点之间的距离，计算方式如下所示：

$$Dis(A,B)=\sqrt{\sum_{i=1}^{n}(A_i-B_i)^2} \tag{7}$$

（三）主题相关性特征

在文本主题关键词的提取方面可以使用隐语义模型（Latent Factor Model）[20]，该算法在文本挖掘领域经常被用到，与之相关的还有 PLSA、LDA 等。

通过获取文本主题关键词，对待消歧文档的 Top N 个词和候选实体集所对应的每篇文档的 Top N 个词进行计算相似度，从而得到所需特征，即候选实体上下文 N 个主题关键词与待消歧实体上下文的 N 个主题关键词相关性的总得分。相关性总得分的计算采用 Google Normalized Distance[21]方式进行统计。

Google Normalized Distance 是基于关系近的概念更有可能出现在同一网页中出现这一假设，然后通过测量两个词语在网页文本中同时出现的频率就可以得到词语间的语义距离。任意两个词 x 和 y，其距离 *GND(x, y)*的计算方式如下所示：

$$GND(x,y)=\frac{G(x,y)-\min(G(x),G(y))}{\max(G(x),G(y))}=\frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log N-\min\{\log f(x),\log f(y)\}} \tag{8}$$

公式中 $f(x)$、$f(y)$分别表示在 Google 中搜 x、y 时对应记录数；$f(x,y)$表示在 Google 中同时搜索 x 和 y 时得到的记录数；N 表示 Google 能检索的 Web 页数总和。$GND(x,y)$表示词语 x 和 y 共现的对称条件概率；假设指定某个页面包含 x（或者 y），则$GND(x,y)$表示该页面同时包含 y（或者 x）的概率，$GND(x,y)$值越大，说明词语 x 和 y 距离越小，语义就越

相关。

4 实验

4.1 实验数据介绍

实验所用数据是由第三届自然语言处理与中文计算会议（简称 NLP&CC2014）提供，NLP&CC2014[6]评测数据的知识库来自中文维基百科中部分含有 InfoBox 结构的实体集。NLP&CC2014 公开的中文微博实体链接评测数据如表 4 所示。

表 4 NLP&CC 2014 中文微博评测数据统计

	训练数据集	测试数据集	测试标注数据集
微博总数	169	1088	570
实体总数	250	1152	607
存在链接的实体	189	—	262
不存在链接的实体（ *NIL* ）	61	—	343

4.2 实验评估指标

本文所采取的度量标准有准确率，召回率以及 F 值，准确率和召回率的计算公式如下所示：

$$Precision = \frac{\left| M \text{ I } M^* \right|}{\left| M \right|} \tag{9}$$

$$Recall = \frac{\left| M \text{ I } M^* \right|}{\left| M^* \right|} \tag{10}$$

式中 M 为实体链接输出的结果， M^* 为标注的正确结果，但是在准确率和召回率往往是相互存在矛盾的，比如为了得到较高的准确率，召回率则会拉低，反之亦然，为了综合考虑准确率和召回率的评价，我们使用 F 值，F 值可以认为是对准确率和召回率的加权调和平均值，公式如下所示：

$$F = \frac{(a^2+1)P*R}{a^2(P+R)} \tag{11}$$

4.3 实验结果分析

本文完成两组对比实验，分别是基于 Lucene 的命名实体链接方法（简称 Lecene_EL）、基于多源知识和 Ranking SVM 的命名实体链接方法（简称 R-SVM_EL）,它们均在 NLP&CC2014 公开的实体链接数据集上进行实验，实验表现的统计有三部分构成，一部分是在整体数据上的准确率，一部分是知识库中存在的目标实体的准确率、召回率及 F1 值，以及知识库中不存在的目标实体的准确率、召回率及 F1 值，而 Best_2014 系统则是 NLP&CC2014 的命名实体链接评测中最佳系统的表现结果,表 5 是系统在整体数据上的结果，表 6 对应知识库中存在的目标实体的结果，表 7 对应知识库中不存在的目标实体的结果。

表 5 NLP&CC2014 整体数据的准确率对比

	Lucene_EL	R-SVM_EL	Best_2014
正确结果总数	372	543	527
待链接实体总数	607	607	607
准确率	0.613	**0.894**	0.868

[6] http://tcci.ccf.org.cn/conference/2014/index.html

表 6 NLP&CC2014 在知识库中存在相应结果的部分的相关结果

	Lucene_EL	R-SVM_EL	Best_2014
准确率	0.5152	**0.8281**	0.8078
召回率	0.3864	**0.8939**	0.8598
F1 值	0.4416	**0.8597**	0.8330

表 7 NLP&CC 2014 在知识库中没有链接的结果

	Lucene_EL	R-SVM_EL	Best_2014
准确率	0.6601	**0.9534**	0.9202
召回率	0.7872	**0.8950**	0.8746
F1 值	0.7181	**0.9233**	0.8969

由于 Lecene_EL 系统是单纯地利用 Lucene 的检索功能，得到与查询名称相似度最高的知识库中的目标实体，因此没有过多复杂的特征和算法。在实验数据集中，由于知识库中无对应词条的待链接实体所占比例基本上为 50%，并且真正存在歧义的待链接实体个数不多，因此在仅利用字符串相似度检索，即基于 Lucene 的实体链接策略的情况下，在 NLP&CC 2014 数据集上取得了相对不高（0.613）的准确率。

从表 6 可以发现，基于 Lucene 的实体链接，在对知识库建立索引库后，进行的字符串匹配检索，从而得到相似度最高的候选实体作为所需的候选实体，由于对相似度得分设置了一定的阈值，并且微博的长度通常比较短，与知识库中的实体信息相比，可能存在语义不充分的情况，导致单纯基于字符串相似度进行比较时，相似度得分有所降低，最终影响召回率；与 Lucene 相比，而 R-SVM_EL 系统则融合了更多的语义特征，比如字符串的表面性特征、主题相关性特征等。

从表 7 中可以发现，由于设定了相似度阈值，因此对于在知识库中不存在链接实体的情况下，取得了较好的召回率及 F1 值，这也充分的证明了阈值设置对于 Lucene 检索策略的重要性，相似度阈值的设置在一定程度上避免了将在知识库中不存在的待链接实体错误地链接到知识库，从而提高了准确率、召回率和 F1 值。

5 结论及下一步工作

本文通过对命名实体链接的研究与实验，分析了课题中的问题，并且提出了相应的研究方法、解决路线及技术框架。本文借助自然语言处理的开源工具和网络百科资源对文本进行数据预处理工作，包括同义词表的构建，同时利用 Lucene 对知识库建立索引。基于 Ranking SVM 模型的命名实体链接，从初始候选实体获取诸多语义特征，利用 RankingSVM 模型对候选实体排序，最终得到最优的目标实体。

通过在 NLP&CC 2014 命名实体链接公开数据集上的对比实验，下一步计划在知识资源和更多深层有效的语义特征方面进行深入的发掘，这对于命名实体链接问题的解决起着关键性的作用。

参考文献

[1] 中国互联网信息中心. 第 35 次中国互联网络发展状况统计报告[R].北京：中国互联网信息中心.2015.1

[2] 郭宇航, 秦兵, 刘挺, 等. 实体链指技术研究进展[J]. 智能计算机与应用, 2014, 4(5).

[3] Mihalcea R, Csomai A. Wikify!: linking documents to encyclopedic knowledge[C]//Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM,

2007: 233-242.
[4] Milne D, Witten I H. Learning to link with wikipedia[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 509-518.
[5] Bunescu R C, Pasca M. Using Encyclopedic Knowledge for Named entity Disambiguation[C]//EACL. 2006, 6: 9-16.
[6] Cucerzan S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data[C]//EMNLP-CoNLL. 2007, 7: 708-716.
[7] Gottipati S, Jiang J. Linking entities to a knowledge base with query expansion[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 804-813.
[8] Sun Y, Zou X, Lin L, et al. ITNLP Entity Linking System at TAC 2013[J].
[9] Zhang W, Sim Y C, Su J, et al. Nus-i2r: Learning a combined system for entity linking[C]//Proc. TAC 2010 Workshop. 2010.
[10] Varma V, Bysani P, Kranthi Reddy V B, et al. iiit hyderabad at tac 2009[C]//Proceedings of Test Analysis Conference 2009 (TAC 09). 2009.
[11] Han X, Zhao J. Nlpr_kbp in tac 2009 kbp track: a two-stage method to entity linking[C]//Proceedings of Test Analysis Conference 2009 (TAC 09). 2009.
[12] Zheng Z, Li F, Huang M, et al. Learning to link entities with knowledge base[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 483-491.
[13] Zhang W, Su J, Tan C L, et al. Entity linking leveraging: automatically generated annotation[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010: 1290-1298.
[14] Han X, Zhao J. Structural semantic relatedness: a knowledge-based method to named entity disambiguation[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 50-59.
[15] Herbrich R, Graepel T, Obermayer K. Large margin rank boundaries for ordinal regression[J]. Advances in neural information processing systems, 1999: 115-132.
[16] Cao Y, Xu J, Liu T Y, et al. Adapting ranking SVM to document retrieval[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 186-193.
[17] Joachims T. Optimizing search engines using clickthrough data[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 133-142.
[18] Joachims T. Training linear SVMs in linear time[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 217-226.
[19] Dill S, Eiron N, Gibson D, et al. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation[C]//Proceedings of the 12th international conference on World Wide Web. ACM, 2003: 178-186.
[20] Chang A X, Spitkovsky V I, Yeh E, et al. Stanford-UBC entity linking at TAC-KBP[J]. Proceedings of TAC, 2010, 758.
[21] McNamee P. HLTCOE efforts in entity linking at TAC KBP 2010[C]//Proc. TAC 2010 Workshop. 2010.

作者简介：

陈万礼（1992——），通讯作者，男，硕士，主要研究领域为自然语言处理。
Email:wanli2013nlp@foxmail.com

咎红英（1966——），女，教授，主要研究领域为自然语言处理。
Email:iehyzan@zzu.edu.cn

吴泳钢（1987——），男，硕士，主要研究领域为自然语言处理。
Email:wygchina@sina.com