北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.042

一个中文实体链接语料库的建设

舒佳根[1,2]　惠浩添[1,2]　钱龙华[1,2,†]　朱巧明[1,2]

1. 苏州大学自然语言处理实验室，苏州 215006; 2. 苏州大学计算机科学与技术学院，苏州 215006;
† 通信作者, E-mail: qianlonghua@suda.edu.cn

摘要　鉴于现有中文实体链接基准语料库的缺乏，在 ACE2005 中文语料库和中文维基百科的基础上，通过自动构造和人工标注的方法，构建了一个中文实体链接语料库及其相关的中文知识库。与传统的英文实体链接语料库不同，构造的中文实体链接语料库是基于实体而非单个实体指称(Mention)。中文实体链接语料库的构建，将为中文实体链接研究提供一个可用的基准平台。

关键词　中文；实体链接；语料库

中图分类号　TP391

Construction of a Chinese Entity Linking Corpus

SHU Jiagen[1,2], HUI Haotian[1,2], QIAN Longhua[1,2,†], ZHU Qiaoming[1,2]

1. Natural Language Processing Lab, Soochow University, Suzhou 215006; 2. School of Computer Science and Technology, Soochow University, Suzhou 215006; † Corresponding author, E-mail: qianlonghua@suda.edu.cn

Abstract　In view of the lack of Chinese entity linking benchmark corpus, the methodology of automatic construction and manual annotation was applied to build a Chinese entity linking corpus as well as its related Chinese knowledge base derived from the ACE2005 Chinese corpus and the Chinese Wikipedia resource. Contrary to traditional English entity linking corpus, this corpus is based on entities rather than individual entity mentions. The construction of Chinese entity linking corpus provides a benchmark platform to the Chinese entity linking research community.

Key words　Chinese; entity linking; corpus

实体链接(entity linking)任务是由 TAC (Text Analysis Conference, http://www.nist.gov/tac)提出的 KBP (knowledge base population)任务中的一个子任务，它的任务就是要将例 1 中的人物实体“刘敬民”链接到知识库中“刘敬民”这个实体上。如果知识库中没有“刘敬民”这个实体，那就返回一个空链接(Nil)[1]。

例 1　[北京奥申委常务副主席] [刘敬民]在接受记者采访时表示，此行达到了[他]预想效果。

其中，“[ ]”内表示实体指称(Mention)。根据 ACE2005 的定义，实体是存在于世界上的某一个对象或者对象的集合。实体的指称是文本中对该实体的引用，一般包括 3 种形式：名称指称(Name Mention)、名词或者名词短语指称(Nominal Mention)、代词指称(Pronoun Mention)[2]。在例 1 中，“刘敬民”是名称指称，“北京奥申委常务副主席”是名词指称，而“他”则是代词指称。

实体链接在信息融合、知识获取以及知识图谱等自然语言处理研究和应用领域有重要的意义，目前的实体链接主要是针对英文的。虽然在 TAC2011 中引入了中文实体链接任务，但它的实质是将中文实体指称链接到英文知识库中，因而属于跨语言实体链接范畴。鉴于中文实体链接语料库的缺乏，本文在 ACE2005 以及中文维基百科的基础

国家自然科学基金(61373096, 90920004)、江苏省高校自然科学重大项目(11KJA520003)资助
收稿日期: 2014–06–29; 修回日期: 2014–10–14; 网络出版时间: 2014-11-28 10:56

上，构建了一个中文实体链接语料库及其相应的中文知识库。

1 相关工作

英文实体链接语料库最早是 2010 年由 TAC 会议发布的。该语料库的标注工作由 LDC 完成，其中包含人物实体(PER)1877 个、组织实体(ORG)3960 个和地理政治实体(GPE)1817 个，语料来源主要包括新闻以及 Web 文本。TAC2010 语料库对 3 种类型(PER, ORG 和 GPE)实体标注的一致性分别为 91.53%, 87.5%和 92.98%; 其知识库来源于 2008 年的英文维基百科，总共包括 818741 个不同类型的实体[3]。

TAC2011 发布的实体连接语料库，在 TAC2010 的基础上增加了跨语言实体连接的语料库。它采用 100 万个来自于中文 Gigaword 语料库的新闻报道作为源文本，包括 1641 个人物实体、1327 个组织实体和 1370 个地理政治实体，知识库采用 TAC2010 的英文实体连接的知识库，将中文实体链接到英文知识库中。TAC2011 的英文实体链接的语料库与 TAC2010 保持一致[4]。

中文方面有 NLPCC2013 发布的中文微博实体链接语料库，该语料库包括 10 个话题，每个话题采集大约 1000 条微博，共约 10000 条微博，平均每条微博包含 1~2 个待测定字符串[5]，所用的知识库基于百度百科[6]。该知识库只包含百度百科中信息盒(Infobox)的信息，且微博文本具有短小精悍的特点，而我们的知识库不仅包含维基百科 Infobox 的信息，还包含相应的文本，而且维基百科信息的结构性更好，语料文本也更具有普遍性。我们的知识库与 TAC 会议 KBP 任务的知识库更为相似。

Cucerzan[7]、Csomai 等[8]、Milne 等[9]和 Kulkarni 等[10]都做过与本次标注类似的任务，但是他们的工作并不局限于实体，还包括词语和短语等。虽然他们的标注工作包含更多的信息，但是标注代价非常大，耗时很长。Bentivogli 等[11]的工作与本文工作最相近，不过他们对每一个实体指称都进行标注，而且链接的维基页面也不是唯一的，语料库的主要目的是用于文本内指代消解的研究。由于本文已经获得文本内的指代信息，所以可以对某一个实体的指代链进行标注，且实体链接的对象是唯一的。

2 语料库的标注

中文实体链接语料的构造分成两个步骤：首先从中文维基百科中构造出中文知识库，然后将 ACE2005 中的中文实体链接到知识库中所对应的实体上去。

2.1 基于中文维基百科的知识库构造

利用中文维基百科 2014 年 1 月份的离线数据包，从中解析出所有中文维基百科的页面。在构建知识库中的实体集合时，本文利用中文维基百科页面内的 Infobox 信息来构建条目的属性信息集合，使用启发式规则的方法来判断维基条目的实体类型，并将页面内的正文作为消岐文本。本文利用中科院分词工具 ICTCLAS2014①对消岐文本进行分词处理。中文知识库中的实体条目如例 2 所示，它是本文知识库收录实体的一个例子，即“王审知”这个人物实体在知识库中的保存形式，其中知识库里实体的 id 就是对应页面在维基百科中的 id, wiki_title 是维基条目的标题。

本文所构造的知识库共收录实体 243520 个，其中人物实体 123318 个、组织实体 27110 个、地理政治实体 78092 个、设施实体(FAC)10880 个、位置实体(LOC)4120 个。

例 2

<entity wiki_title=“王审知” type=“PER” id=“E205886” name=“王审知”>
<facts class=“Emperorcn box”>
<fact name=“姓名”>王审知</fact>
……
</facts>
<wiki_text><![CDATA[王审知
闽太祖王审知，字信通，一字详卿。光州固始(今河南固始)人……
]]></wiki_text>
</entity>

2.2 ACE2005 中文语料库

中文 ACE2005 语料库由 633 篇来自广播、新闻报道、网络博客、转录音频等不同领域的文本组成，总共包含 6771 个实体，实体类型有 PER, ORG, GPE, LOC, FAC, WEA(武器实体)和 VEH(交通工具

① http://ictclas.nlpir.org/newsDetail?DocId=390

实体)等 7 类。ACE2005 除标注实体及其指称外, 还包括文本内的实体指代链, 所以在本文的标注过程中只要对指代链中任意一个指称(Mention)进行标注, 整个实体链就标注好了, 这就加快了标注的速度。

2.3 标注方法及过程

为了提高标注速度和可靠性, 基于 ACE2005 中文语料库的实体连接语料库的标注过程分为自动标注和人工校正两个过程。

2.3.1 自动标注

首先, 本文把某一实体的指代链中的最长指称和维基中的实体名称(对应实体条目的标题, 包括重定向页面以及消岐页面的标题)进行比对, 当它们完全匹配时进行自动标注, 即将该实体链接到知识库中对应的实体条目中。

2.3.2 人工校正

由于实体重名和多名现象的存在, 自动标注的结果中存在很多错误, 需要进行手工调整。有些实体的指称由于找不到精确匹配的对象, 无法进行自动标注, 也需要手工标注。

例 3 “原中国纺织总会会长、第九届全国政协常委[吴文英]因严重违纪……”

例 4 “南宋诗人[吴文英]本姓翁, 后来过继给姓吴的人改姓吴。”

例 5 “ [桃园机场]是当时亚洲最现代化的国际机场之一。”

例 6 “行政院院会决议为纪念中华民国已故总统蒋中正而命名为[中正国际机场]。”

其中例 3 和例 4 两个句子是重名现象的一个例子, 其中的两个人物虽然名称相同, 但他们不是同一个人。例 5 和例 6 两个句子是多名现象的一个例子, 两句中的机场名字虽然不同, 但它们指向同一个机场。

图 1 是针对本次标注任务开发的标注工具, 主要由 4 部分组成: 左栏用来选择标注文件; 右栏是该文件所对应的文本, 是标注主要进行的地方; 顶栏是本软件的工具栏; 底栏是当前选中实体的一些信息。在右栏中分别用 5 种不同颜色来表示我们所需要标注的 5 种实体类型。对某一实体进行链接标注时, 将鼠标移到该实体的某一指称上, 选择右键菜单中的“Link to Wiki”菜单项, 输入该实体的维基 id (实体的 id 可从图 2 软件的右栏得到), 再点击 SaveLink, 就可以完成对该实体的链接标注。如果自动标注的结果正确, 则无需任何调整。图 2 是用于显示中文维基百科中所有条目的软件工具, 点击左栏中的条目, 右栏中就会显示该条目所对应的维基文本。该软件也可以用维基条目的名称进行反向搜索。

本文的标注主要利用实体指称以及它所在的源文档信息进行标注, 如图 1 中的“马英久”这个实体, 仅通过字符串精确匹配找不到对应实体, 所以必须利用文本中与“马英久”相关的信息, 发现“马英久”是台北市长, 而本文档的来源是 2000 年的 CBS (台湾“中央广播电台”)新闻, 所以可以通过查找 2000 年的台北市长发现知识库中的“马英久”。图 2 是维基中对“马英久”的介绍, 我们可以确认它就是“马

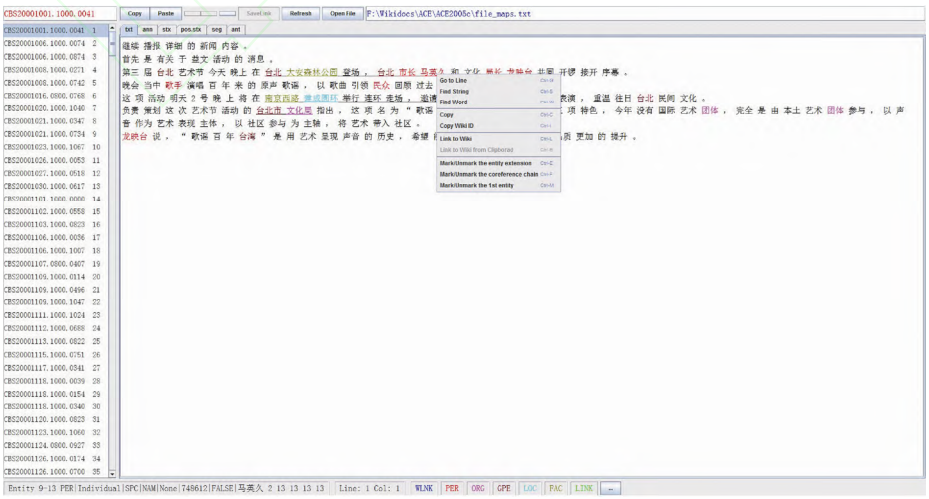

图 1 标注工具界面

Fig. 1 Interface of annotation software

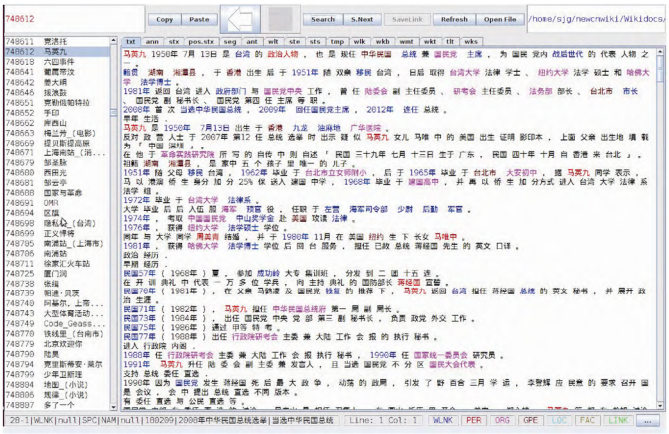

图 2　维基条目显示界面

Fig. 2　Interface of wikipedia display software

英久”这个实体指称在知识库中对应的实体。

2.4　语料库统计与分析

本节将对标注得到的实体连接语料库从实体类型分布、中文维基百科覆盖率和标注一致性 3 个不同的角度进行分析。

2.4.1　实体类型分布情况

本文标注的实体链接语料库包括人物、组织、地理政治实体、位置实体和设施实体等 5 种实体类型，指称分布情况如表 1 所示。其中，地理政治实体的比例最高，达到 35.8%；组织实体与人物实体基本持平，都在 28%左右；位置实体与设施实体相对来说较少，均不足 5%。

2.4.2　中文维基百科对 ACE 实体的覆盖率

如果 ACE2005 中的某个实体可以在知识库中找到相应的实体，就认为维基百科覆盖该实体。实体覆盖率指文本中的实体被维基百科覆盖的比率，计算方法为

$$\text{覆盖率}=\frac{\text{覆盖的实体总数}}{\text{所有的实体总数}}*100\%\ 。$$

表 1　ACE2005 中不同类型的实体指称分布

Table 1　Proportions of different entity type in ACE2005

实体类型	指称数目	所占比例/%
PER	1860	27.6
ORG	1939	28.7
GPE	2419	35.8
LOC	233	3.5
FAC	299	4.4
合计	6750	100

表 2 比较了维基百科 5 种实体类型的覆盖率，可以发现总体覆盖率超过 70%。当然，不同类型实体的覆盖率不同，其中地理政治实体的覆盖率最高(97%)，而人物实体的覆盖率较低，不足 45%，主要原因是 ACE2005 文档中有众多的不知名人物，而维基百科对不知名人物一般没有对应的条目，因此造成人物实体覆盖率低下。组织实体的覆盖率也只有 66%左右，这也是由于一些不太知名的组织在维基百科中没有对应的词条。地理政治实体的覆盖率较高，表明 ACE2005 中绝大多数的地理政治实体都是知名的，因而收录在维基百科中。

2.4.3　一致性检验

本文引入 Coincidence 度量值[5]，即通过计算两位标注者之间的一致性程度来衡量语料标注的可靠性，计算方法为

$$\text{Coincidence}=\frac{2|(S_1\cap S_2)|}{|S_1|+|S_2|}\ ,$$

其中$|S_1\cap S_2|$表示标注者 1 和标注者 2 的标注

表 2　维基百科对 ACE2005 实体指称的覆盖率

Table 2　Coverage of Wikipedia for the entity mentions in ACE2005

实体类型	覆盖实体数目	实体覆盖率/%
PER	815	43.8
ORG	1349	65.6
GPE	2350	97.0
FAC	170	56.9
LOC	181	77.7
合计	4995	72.1

结果中一致的实体数目, $|S_1|$代表标注者 1 标注的实体总数, $|S_2|$代表标注者 2 标注的实体总数。Coincidence 值越高, 说明一致性越好; 反之, 则一致性越差。

从 633 个文档中随机抽取 105 个文档作为一致性检查的数据集, 并对这些文档进行一次自动标注和两次独立的人工标注。第一次人工标注由两个初级标注者 A 和 B 完成, 他们没有 NLP (自然语言处理)学科背景, 只经过初步培训。第二次人工标注由两位具有 NLP 学科背景的标注者 C 和 D 完成。最后我们请第 5 个标注者 G 对 C 和 D 中不一致的地方再进行调整, 完成最终数据集的标注。

1) 自动标注与最终标注结果的一致性情况。

表 3 列出自动标注和最终标注之间的一致性结果, 其中 T 为自动标注集, Z 为最终的语料标注集。

从表 3 可以看出, FAC 和 LOC 的自动标注结果还是不错的, 与最终结果的一致性达到 73%左右, 主要是这两种实体的重名以及多名现象比较少, 因而对这两种实体的实体链接任务相比其他类型的实体要简单一些。但是, 人物实体不到 65%, 组织实体、地理政治实体以及整体不到 60%的一致性, 表明仅靠精确匹配无法较好地完成对语料库的标注工作。

2) 人工标注者之间的一致性情况。

由于 A 和 B 不具备 NLP 学科背景, 只经过简单培训, 所以他们的标注结果必然存在不少错误, 其标注的一致性结果见表 4, 其中 A, B, C, D 和 G 分别表示相应标注者所标注的数据集。

从表 4 可以看出, 该语料库经过二次标注之后, 标注的一致性有极大的提升, 达到 97%, 甚至超过 TAC 发布的英文实体链接语料库的一致性, 说明此时语料库标注的可信度是相当高的。

表 3 自动标注集和最终标注集的一致性

Table 3 Consistent rate of automatic annotation set and final annotation set

实体类型	Coincidence/%
	T 和 Z
PER	63.9
ORG	58.5
GPE	55.1
FAC	72.8
LOC	73.0
合计	58.8

表 4 调整前后的一致性

Table 4 Coincidence scores before and after adjustment

实体类型	Coincidence/%			
	A 和 B	C 和 D	C 和 G	D 和 G
PER	74.0	97.5	99.5	98.0
ORG	67.7	95.2	98.4	96.7
GPE	74.1	97.9	98.9	99.5
FAC	72.4	95.8	100	100
LOC	70.8	96.6	91.7	100
ALL	72.0	97.0	98.8	98.5

3) 语料标注的不一致性分析。

在表 3 和 4 的基础上进一步分析不一致性, 如表 5 所示, 分别在调整前和调整后两个数据集上进行。

从表 5 可以发现, A 和 B 标注不一致的实体类型的分布情况, 与整体的实体类型分布情况基本一致, 可以认为每种类型出现不一致的概率基本一样。C 和 D 的不一致实体类型分布与整体的实体

表 5 不一致分析

Table 5 Analysis of inconsistency

实体类型	A 和 B 不一致		C 和 D 不一致		105 个文档实体分布	
	数目	比例/%	数目	比例/%	数目	比例/%
PER	52	31.3	5	26.2	200	33.73
ORG	40	24.1	6	31.6	124	20.1
GPE	50	30.1	4	21.1	187	31.5
FAC	17	10.2	1	5.3	58	9.8
LOC	7	4.3	3	15.8	24	4.3
合计	166	100	19	100	593	100

类型分布并不一致, 表明对于不同类型的实体, 调整的效果不一样。对于 PER 和 GPE 的调整效果明显优于其他 3 种实体类型, ORG 类型的实体链接相对较难。

通过对 A 和 B, 以及 C 和 D 不一致实体的调研, 发现 A 和 B 一致性较差的原因是, 他们没有很好地针对语言现象(即实体的重名以及多名现象)进行处理, 造成很多不一致; 而 C 和 D 由于具有 NLP 学科背景, 能对实体的多名和重名现象进行处理, 因而提高了一致性。

3 语料库上的基准系统

3.1 实验方法

实验方法分 3 步进行, 本文基准系统的流程如图 3 所示。

1) 候选集生成。通过计算实体指称与知识库中实体名字之间的字符串相似度来生成候选集, 即当知识库中某个实体的名字与实体指称的相似度大于某个阈值时(本文取 0.5), 将该实体加入候选集[12]。本文采用编辑距离来计算实体指称与知识库中的实体名字之间的相似度。同时, 为了尽可能使候选集包含正确答案, 我们抽取维基百科的重定向列表, 只要实体指称与知识库中某个实体重定向列表中的任意一个名字的相似度大于 0.5, 就将该实体加入候选集, 并且字符串相似度的值取重定向列表中相似度最高的那个值。

2) 实体消岐。采用无监督排序的方法, 将字符串相似度和文本相似度加权的方法来计算总体相似度。其中文本相似度指实体所在文本与候选实体的消岐文本之间的相似度, 采用频数加权的词包(TF-BOW)模型来计算。

3) 空链接(Nil)判断。如果总体相似度的最高值大于我们设置的阈值, 就认为该实体是所要链接的实体, 否则就返回空链接(Nil)。

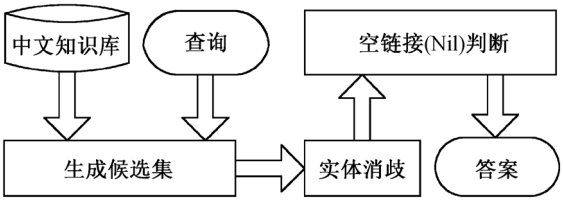

图 3 基准系统流程

Fig. 3 Flowchart of the baseline system

3.2 实验设置

在已经标注好的中文实体链接语料库的基础上, 采用五倍交叉的实验方法, 基准系统采用字符串相似度和文本相似度两个特征进行实体链接。

3.3 实验性能

本文使用精度值 Accuracy 来评估实验系统的性能, 即正确链接的实体数量占全部实体的比例, 精度值越高则系统的性能越好, 其计算方法为

$$\text{Accuracy}=\frac{\text{right\_answers}}{\text{all\_queries}},$$

其中 right_answers 是系统链接正确的实体数(包括正确判断的空链接数目), all_queries 是所有测试数据的实体数。

3.4 实验结果分析

表 6 比较了不同实验方法的结果, 我们发现单独使用字符串相似度这个特征的精度还是可以的, 达到 63.7%; 而单独使用文本相似度特征的性能非常差, 只有 35.4%, 其中 GPE 类型甚至只有 5.9%; 当这两种特征线性按 9:1 加权和归一后, 达到 69.7%。

3.5 错误分析

为了进一步了解错误链接的原因, 我们随机调研了 100 个错误结果, 发现原因有 3 种。

1) 命名实体的多名、不同音译、 错别字等现象: 它造成候选集无法包含正确答案, 正确答案在

表 6 不同方法的实验性能

Table 6 Performance for different methods

实验方案	Accuracy/%					
	PER	ORG	GPE	FAC	LOC	合计
字符串精确匹配(阈值为 1.0)	64.9	60.3	55.1	72.8	73.0	60.8
字符串相似度(阈值为 0.85)	67.1	60.3	59.5	74.1	73.0	63.7
文本相似度(阈值为 0.85)	60.6	36.8	5.9	48.1	47.4	35.4
字符串+文本相似(阈值为 0.85)	69.2	63.2	72.1	79.0	73.0	69.7

候选集中但被判定为空链接等两种错误现象, 占所有错误的 59%。

2) 命名实体重名现象: 它造成系统的消岐错误, 占所有错误的 37%。

3) 命名实体所在文档没有包含该实体的相关信息: 由于文本相似度特征在消岐时采用的比重较小, 所以它造成的消岐错误仅占所有错误的 4%。

4 总结与展望

本文介绍了一个中文实体连接语料库的构建过程, 并分析了语料库的组成和标注的一致性情况。分析表明, 经过调整后的语料库其标注一致性较好, 满足基准语料库的需求。在本文语料库上, 我们还实现了一个实体链接的基准系统。根据对错误原因的分析, 下一步工作需要在基准系统的基础上, 通过改进候选集生成方法, 挖掘更多的语言特征来帮助系统进行实体消岐以及改进空链接(Nil)判定方法, 进一步提高实体链接系统的性能。

参考文献

[1] McNamee P, Simpson H, Dang H T, et al. Overview of the TAC 2009 knowledge base population track.

[2] ACE (Automatic Content Extraction). Chinese Annotation Guidelines for Entities Version 5.5. (2005–05–05) http://www.ldc.upenn.edu/Projects/ACE/

[3] Ji H, Grishman R, Dang H T, et al. Overview of the TAC 2010 knowledge base population track // Third Text Analysis Conference (TAC 2010). Gaithersburg, MD, 2010

[4] Ji H, Grishman R, Dang H T, et al. Overview of the TAC 2011 knowledge base population track // Fourth Text Analysis Conference (TAC 2011). Gaithersburg, MD, 2011

[5] TCCI. 中文微博实体链接评测大纲[R/OL]. (2013) http://tcci.ccf.org.cn/conference/2013/dldoc/ev04.pdf

[6] 朱敏, 贾真, 左玲, 等. 中文微博实体链接研究. 北京大学学报: 自然科学版, 2014, 50(1): 73–78

[7] Cucerzan S. Large-scale named entity disambiguation based on Wikipedia data // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Prague, 2007: 708–716

[8] Csomai A, Mihalcea R. Linking documents to encyclopedic knowledge. IEEE Intelligent Systems, 2008, 23(5): 34–41

[9] Milne D, Witten I H. Learning to link with Wikipedia // CIKM'08: Proceeding of the 17th ACM conference on Information and knowledge management. Hong Kong, 2008: 509–518

[10] Kulkarni S, Singh A, Ramakrishnan G, et al. Collective annotation of wikipedia entities in web text // KDD'09: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Paris, 2009: 457–466

[11] Bentivogli L, Forner P, Giuliano C, et al. Extending English ACE 2005 corpus annotation with ground-truth links to Wikipedia // 23rd International Conference on Computational Linguistics. Beijing, 2010: 19–26

[12] Rao D, McNamee P, Dredze M. Entity linking: finding extracted entities in a knowledge base // Multi-source, Multilingual Information Extraction and Summarization. Berlin: Springer, 2013: 93–115