文章编号：1003-0077(2017)02-0036-06

基于特征加权重叠度的中文实体协同消歧方法

线岩团，余正涛，洪旭东，张 磊，郭剑毅

（昆明理工大学 信息工程与自动化学院，云南 昆明 650500）

摘 要：该文针对中文实体消歧中的特征项部分匹配和协同消歧问题，提出基于特征加权重叠度的中文实体协同消歧方法。该方法利用实体指称上下文中多种特征的加权重叠度计算实体指称相似度，针对实体链接与消歧聚类约束，分类定义实体指称相似度计算方法，构建待消歧实体相似度矩阵，采用近邻传播聚类算法实现中文实体协同链接与消歧。基于 CLP-2012 评测数据的实验表明，提出的方法取得了较好的消歧效果，准确率、召回率和 F 值分别达到了 84.01%、87.75%和 85.65%。

关键词：实体消歧；实体链接；加权重叠度；近邻传播聚类

中图分类号：TP391 **文献标识码**：A

Collaborative Entity Disambiguation Method Based on Weighted Feature Overlap Relatedness for Chinese

XIAN Yantuan, YU Zhengtao, HONG Xudong, ZHANG Lei, GUO Jianyi

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: A collaborative entity disambiguation method based on weighted feature overlap relatedness is proposed in this paper. This method make use of weighted feature overlap relatedness for computing the similarity between entity names. We define some deferent similarity formulas for computing entity similarity matrix, then the affinity propagation clustering algorithm is used to get the disambiguation results. Evaluation on the CLP-2012 corpus shows that our method can achieve competitive performance, attains 84.01% precision, 87.75% recall and 85.65% F-score.

Key words: entity disambiguation; entity linking; weighted overlap relatedness; affinity propagation clustering

1 引言

命名实体识别与消歧是自然语言处理的重要研究点，其主要任务是发现文本中的命名实体，并将实体指称关联到已有真实实体。命名实体识别与消歧已成为知识库构建、信息抽取、机器翻译，以及话题发现与追踪等研究领域的重要支撑技术[1]。

命名实体普遍存在重名和歧义现象，即同一实体指称在不同上下文中可对应不同的实体，例如，“高峰”一词可表示普通的名词，也可能表示不同的人名、机构名，或是地名。针对命名实体消歧已有许多相关的研究，在英文评测方面主要有 TAC KBP 的 Entity Linking 评测[2-3]和 WePS(Web People Search)评测[4]，以及针对中文的 CLP-2012 汉语命名实体识别与歧义消解[5]和 NLP &CC 中文微博实体链接评测。

实体消歧可利用的知识主要有两类，一是实体指称的上下文信息，如实体指称周围的词语、实体等；二是外部知识库，如 Wikipedia、百度百科、DBpedia[6]和 Freebase[7]等。实体消歧方法的核心是计算实体指称和目标实体之间的相似度，并根据实体相似度实现消歧和链接。主要方法有：基于文本向量空间的聚类方法、基于分类的方法和基于图的实体消歧方法。基于文本向量空间的聚类方法通过实体指称的上下文信息构建文本向量，计算文本向

收稿日期：2015-03-10 定稿日期：2015-09-10

基金项目：国家自然科学基金(61363044, 61175068, 61365010, 61462054, 61462055)

量间的距离，最后利用该距离进行聚类，确定实体指称对应的实体概念[8-11]。该类方法在计算相似度时，通常不考虑特征项部分匹配的问题，对相似度计算造成了不利的影响。基于分类的方法利用上下文特征构建分类器，将待消歧实体链接到知识库[12-14]。该方法存在的主要问题是分类的确定和待消歧的数据紧密相关，难以构建通用的分类模型适应不同的消歧问题。Peng 等人提出了结合分类和聚类算法的实体链接消歧方法[15]。该方法利用分类算法将待消歧实体划分为知识库实体、未知实体和普通词三类，然后，选取命名实体、职业和名词等特征构建特征向量，计算特征向量余弦相似度，采用分类方法完成实体链接，最后，利用合成聚类（Agglomerative Hierarchical Clustering，AHC）方法完成未知实体消歧，取得了较好的效果。基于图的实体消歧方法将实体指称看作图中的节点，利用实体指称间的关联关系（链接、实体共现等）构建实体关联图，通过图聚类算法或是阈值过滤得到实体消歧结果[16-18]。该类方法对于具有较多关联关系的实体消歧问题有较好的效果，但对于缺乏关联关系的实体消歧问题，难以构建有效的实体关联图。

目前，已有的实体消歧方法通常依据特征项的共现信息，在计算实体相似度时忽视了特征项部分匹配的问题。针对这一问题，Hoffart 等人提出了基于关键短语集合重叠度的实体指称相似度计算方法[19]。Ikeda 等人提出了利用实体、组合关键词、链接等特征，通过重叠系数计算各类特征的相似度，最后线性加权方式获取人名实体间的关联度[20]。另一方面已有的方法大多将实体链接和未知实体消歧分作两个步骤处理，忽略了多个实体指称之间关联关系对实体消歧的影响。本文针对以上两方面的问题，提出基于特征加权重叠度的中文实体协同消歧方法。该方法利用实体指称上下文中的多种特征的加权重叠度计算实体相似度，针对实体链接与消歧聚类约束，分类定义实体指称相似度计算方法，采用近邻传播聚类算法实现中文实体协同链接与消歧。

2　基于特征加权重叠度的中文实体协同消歧方法

2.1　实体消歧问题定义

命名实体消歧的任务就是利用实体指称上下文和其它相关知识，将待消歧实体指称划分到其对应的真实实体的过程。也就是给定目标实体知识库 $E=\{e_1,e_2,\cdots,e_m\}$ 和待消歧名字集合 $N=\{n_1,n_2,\cdots,n_h\}$，将 N 划分为 k 个不相交的簇集合 $C=\{c_1,c_2,\cdots,c_k\}$，则聚类 $c_i=\{n_p^i,\cdots,n_q^i\}(1\leqslant p\leqslant q\leqslant h)$ 中的实体表示同一实体概念，并将 c_i 链接到 E 中对应的实体概念。

2.2　特征选择与提取

命名实体能够简洁的表示与待消歧实体关联的真实概念，Elmacioglu 等人的研究表明上下文中的命名实体是有效的消歧特征[21]。WePS 系列评测中大部分排名靠前的系统都采用命名实作为主要特征，也从另一个侧面验证了这一点。但是只用命名实体作为消歧特征是远远不够的，主要原因有两个方面。一是待消歧实体上下文中不一定出现命名实体，二是命名实体识别准确度难以保证。本文在分词、词性标注和实体识别基础上，结合职业和职称词典，获取专有名词、人名、机构名、地名、作品名称、职业和职称作为主要的消歧特征。此外，上下文中的概念通常以名词词组形式出现，描述了待消歧实体的属性和特点，能很好的表征实体间的语义关系；所以，本文还选取与前述特征不重叠的名词词组作为特征，计算实体间的语义关联度。获取到的特征项集合由 $\{\mathbf{w}\}$ 表示，特征项 $\mathbf{w}=\{w_1,w_2,\cdots,w_i,\cdots,w_L\}$，其中 w_i 是特征项中的词语，L 是特征项的长度。

2.2.1　特征项重叠相似度计算

命名实体和名词词组特征通常由多个词语构成，所以，这些特征项之间的部分匹配就显得十分重要，例如“退役足球运动员 ”和“足球运动员”的相似度应高于它和“蓝球运动员”的相似度。为了解决这一问题，本文借鉴 Hoffart 等人提出的短语相似度的计算方法[19] 分别计算命名实体和名词词组之间的重叠相似度。

设 (e,f) 表示待消歧的实体对，$P_e=\{p_1,p_2,\cdots\}$，$P_f=\{q_1,q_2,\cdots\}$ 分别表示实体 e 和 f 的特征项集合。特征项由词语构成，即 $p_i=\{w_1,w_2,\cdots\}$，w_i 具有相对于实体 e 的权重 $\gamma_e(w_i)$。为了简化公式，文本用 p 和 q 表示不同实体的特征项，w 表示特征项中的词，则基于加权 Jaccard 相似性系数的特征项重叠相似度如式(1)所示。

$$\mathrm{po}(p,q)=\frac{\sum\limits_{w\in p\cap q}\min\{\gamma_e(w),\gamma_f(w)\}}{\sum\limits_{w\in p\cup q}\max\{\gamma_e(w),\gamma_f(w)\}} \tag{1}$$

特征项词语权重 $\gamma_e(w)$ 和 $\gamma_f(w)$ 由实体上下文中的逆文档频率(IDF,Inverse Document Frequency)确定,即 $\gamma(w)=\log_2(Z/df(w))$,$Z$ 表示待消歧实体和目标实体的总数,$df(w)$ 表示包含 w 的特征项出现的次数。

在特征项重叠相似度 $\mathrm{po}(p,q)$ 基础上,计算实体对 (e,f) 的关联度如式(2)所示。

$$\mathrm{sim}(e,f)=\frac{\sum_{p\in P_e,q\in P_f}\mathrm{po}(p,q)\times\min\{\varphi_e(p),\varphi_f(q)\}\times\min\{\psi(p),\psi(q)\}}{\sum_{p\in P_e}\varphi_e(p)+\sum_{q\in P_f}\varphi_f(q)} \tag{2}$$

其中 $\varphi_e(p)$ 表示特征项 p 相对于实体 e 的权重,由 p 的 TF-IDF 值确定。$\psi(p)$ 表示不同类型的特征的权重。式中分子选择较小的特征项权重和特征类型权重,对 $\mathrm{po}(p,q)$ 再加权。分母通过实体特征项权重之和,对 $\mathrm{sim}(e,f)$ 进行正规化。在此,分母不采用特征项交集的最大权值求和,以避免对特征项的迪卡尔集进行计算,降低计算复杂度。考虑到不同类型的特征对实体消歧结果的贡献存在差异,本文针对不同类别特征类型定义不同的权重。权重值基于少量数据采用最小错误率训练算法[22]得到,得到的具体仅值参见表1,权重调优公式参见式(3)。

$$\hat{\lambda}_1^M=\underset{\hat{\lambda}_1^M}{\mathrm{argmin}}\left\{\sum_{i=1}^{Z}E(r_i,\hat{r}(n_i;\lambda_1^M))\right\} \tag{3}$$

其中,$\hat{\lambda}_1^M$ 表示待调整的 M 个权值,$\hat{r}(n_i;\lambda_1^M)$ 表示待消歧名字实体在参数 $\hat{\lambda}_1^M$ 条件下的消歧结果,$E(r_i,\hat{r}(n_i;\lambda_1^M))$ 是消歧结果和答案 r_i 相比存在的错误的个数。Z 表示用于训练的名字实体数。

表1 特征类型权重

特征类别	特征类别标记	权重值
人名	PER	0.48
机构名	ORG	0.37
地名	LOC	0.63
职业、职称名	OCP	0.95
专有名词	NZ	0.28
名词词组	PHR	0.35
作品名	OPUS	0.71

2.3 基于近邻传播的协同实体链接消歧

本文基于近邻传播聚类算法[23](Affinity Propagation cluster algorithm,AP)实现协同实体链接与消歧。实体链接与消歧可看作是针对实体知识库 E 和待消歧名字集合 N 的聚类问题,针对实体链接与消歧任务要求,聚类应满足以下的约束条件:

a. 由于待消歧名字集合对应的真实概念是不确定的,故聚类数目 k 也是不确定;

b. 实体知识库 E 中的实体表示独立的含义,所以它们之间的关联度为零;

c. 实体知识库 E 中的实体通常包含更准确、更丰富的上下文信息,应具有更大机会成为聚类代表结点(Exemplar);

d. 聚类过程应协同考虑多个待消歧实体间的关联度和相互作用,从而更好地对多个实体进行消歧。

近邻传播聚类算法是一种基于近邻信息传播的聚类算法,其目的是找到最优的类代表点集合,使得所有数据点到最近的类代表点的相似度之和最大。本文将实体链接与消歧看作针对待消歧名字集合 N 与实体知识库 E 的近邻传播聚类问题。其中,聚类结点集合 $V=\{v_i\mid v_i\in E\cup N\}$,结点间关联度矩阵 S 按式(4)计算得到。

$$\begin{cases}s_{ij}=sim(v_i,v_j) & \text{if } i\neq j,v_i\in N,v_j\in V\\ s_{ij}=0 & \text{if } i\neq j,v_i\in E,v_j\in E\\ s_{ii}=Q_{\frac{1}{2}}(\mathrm{S}) & \text{if } v_i\in N\\ s_{jj}=\alpha Q_{\frac{1}{2}}(S) & \text{if } v_j\in E\end{cases} \tag{4}$$

式中 $\mathrm{sim}(v_i,v_j)$ 表示由特征加权重叠相似度表征的结点相似度,$Q_{\frac{1}{2}}(\mathrm{S})$ 表示关联度矩阵 S 去除对角线元素后的中位数。以上关联度矩阵计算方法很好地满足了协同实体链接消歧的约束条件,其中约束 a 和 d 由近邻传播聚类算法本身特性来满足;约束 b 由公式(4)中的 $s_{ij}=0$ 实现,公式(4)中的 $s_{ii}=Q_{\frac{1}{2}}(\mathrm{S})$ 表示待消歧实体成为聚类代表点机率。$s_{jj}=\alpha Q_{\frac{1}{2}}(\mathrm{S})$ 中的系数 α 用于提高知识库实体成为聚类代表点的机会,在此我们根据实验选取经验值 $\alpha=1.25$。

3 实验结果及分析

本文采用 CLP-2012"汉语命名实体识别与歧义消解"评测任务提供的训练数据开展实验。该评测

任务提供的数据包含 16 个实体指称，每个实体指称对应 50～200 篇文本。

CLP-2012 评测提供的实体知识库包含实体指称 Name 的多个不同定义，每个定义由一段文字描述。针对每个实体指称 Name，有一个包含 Name 的文本集合 T。要求判断 T 中的实体指称 Name 是实体名，还是普通词。如果 Name 是实体名，则进一步判断它是否对应于知识库中的定义；如果 Name 未对应到知识库中的定义，则将按其含义将实体指进行归类。如 Name 是普通词，则将其归入 Other 类中。

为了避免 Other 类的判别问题，实验过程中借助于 Peng 等人提出的方法[15]，将 Name 作为检索词，利用互联网搜索引擎获检索结果中的前 50 个页面，获取与 Name 共现度最高的 20 个名词词组。利用获取的词组在实体知识库中加入与 Other 类对应的伪实体定义，则 Other 类的判别转化为在实体知识库寻找对应定义的问题。词语与 Name 共现度按式(5)计算。

$$co(name,word)=\frac{d(name,word)}{d(name)+d(word)} \quad (5)$$

式中，$d(name,word)$ 表示同时包含 Name 和 Word 的文档数，$d(name)$ 表示包含 Name 的文档数量，$d(word)$ 表示包含 Word 的文档数。

本文将实体链接与消歧看作是以实体指称为结点的聚类问题，聚类结果中同一个聚类簇的实体指称具有相同含义。实验过程主要包含以下三个步骤。首先利用词法分析工具[24]对知识库中的实体定义和文本集合 T 进行分词、词性标注和命名实体识别，并抽取实体特征和名词词组特征。然后，利用特征值计算结点似度矩阵。其中，不同实体指称间的相似度基于特征加权重叠度方法计算，即公式(1)和(2)；其它情况按公式(4)进行计算。最后，采用近邻传播聚类工具[25]对实体指称进行聚类，从而得到实体链接与消歧结果。

实验结果的准确率和召回率按 CLP-2012 评测任务提供的公式计算，实验结果见表 2。

表 2　CLP-2012 训练数据实验结果

待消歧实体名	准确率	召回率	F 值
白雪	0.816 6	0.854 7	0.835 2
白云	0.638 6	0.846 2	0.727 9
丛林	0.923 8	0.845 9	0.883 1
杜鹃	0.864 4	0.850 9	0.857 6
方正	0.894 1	0.831 1	0.861 4
高超	0.823 6	0.906 7	0.863 2
高峰	0.690 3	0.794 3	0.738 7
高明	0.888 3	0.911 7	0.899 8
高山	0.905 6	0.920 4	0.912 9
高雄	0.916 8	0.910 2	0.913 5
胡琴	0.945 7	0.926 1	0.935 8
华明	0.970 5	0.926 4	0.947 9
华山	0.760 4	0.819 3	0.788 8
黄海	0.856 1	0.904 5	0.879 6
黄河	0.694 1	0.866 7	0.770 9
雷雨	0.851 9	0.925 3	0.887 1
total	**0.840 1**	**0.877 5**	**0.856 5**

实验结果表明本文提出的方法在准确率、召回率和 F 值上取得了较好的效果。实验结果与 CLP-2012 评测结果比较参见表 3。实验结果与 Peng 等人提出的两阶段实体消歧方法[15]结果相近。该方法首先利用分类算法将待消歧实体指称划分已有实体、未知实体和普通词三类；然后，利用层次聚类算法对未知实体进行消歧。相比于前述的两阶段实体消歧方法，本文方法只采用近邻传播聚类算法实现消歧，在简化方法的同时实现了实体指称的协同消歧。

表 3　实体消歧方法结果比较

实体消歧方法	准确率	召回率	F 值
UM	0.744 1	0.690 9	0.716 5
UM-1	0.685 9	0.860 0	0.763 2
ZZU	0.639 9	0.679 5	0.659 0
HITSZ	0.676 1	0.727 7	0.701 0
BUPT	0.671 8	0.856 2	0.752 9
SIR-NERD	0.833 2	0.879 0	0.855 5
KMUST-LIIP	0.840 1	0.877 5	0.856 5

注：KMUST-LIIP 表示本文提出的方法，UM 和 UM-1 表示来自文献[12,26]针对部分训练数据得到的结果，ZZU、HITSZ、BUPT 和 SIR-NERD 结果分别来自文献[10-15]。

通过对实验过程分析发现，命名实体识别工具针对 CLP-2012 评测数据中的实体识别准确率低是

影响实验结果的主要因素之一。究其原因是评测试数据中出现的许多命名实体和普通词语重名，导致了较多的识别错误，最终影响了实验结果的准确度。另外，职业、职称和作品名等特有特征对人名实体消歧具有很好的效果，其平均准确率、召回率和F值高于其它的实体消歧结果7.2%，4.0%和5.7%。

4 结论

本文针对中文实体消歧中的特征项部分匹配的问题，提出基于特征加权重叠度的中文实体协同消歧方法。实验结果表明该方法针对面向文本中文实体链接与消歧问题具有较好的效果。本文利用实体指称上下文中的多种特征的加权重叠度计算实体相似度，能较好具体现实体间的语义关联，但由于待消歧实指称上下文信息量有限，而外部知识库（Wikipedia、百度百科和互动百科等）含有大量的实体信息，所以，我们将进一步研究融合外部知识库语义相似度计算方法，提高中文实体消歧的性能。

参考文献

[1] 赵军. 命名实体识别，排歧和跨语言关联[J]. 中文信息学报，2009，23(2)：3-17.

[2] Ji H，Grishman R. Knowledge base population：Successful approaches and challenges[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics：Human Language Technologies-Volume 1. Association for Computational Linguistics，2011：1148-1158.

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

[4] Artiles J，Gonzalo J，Sekine S. The semeval-2007 weps evaluation：Establishing a benchmark for the web people search task[C]//Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics，2007：64-69.

[5] Wang Z H H，Li S. The Task 2 of CIPS-SIGHAN 2012 Named Entity Recognition and Disambiguation in Chinese Bakeoff[C]//Proceedings of The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing（CLP-2012）. 2012：108-114.

[6] Auer S，Bizer C，Kobilarov G，et al. Dbpedia：A nucleus for a web of open data[M]//The semantic web. Springer Berlin Heidelberg，2007：722-735.

[7] Bollacker K，Evans C，Paritosh P，et al. Freebase：a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM，2008：1247-1250.

[8] Cucerzan S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data[C]//Proceedings of the EMNLP-CoNLL. 2007，7：708-716.

[9] 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.

[10] Fan Q，ZAN H，CHAI Y，et al. Chinese personal name disambiguation based on vector space model[C]//Proceedings of The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing（CLP-2012）. 2012：152-158.

[11] Cilibrasi R L，Vitanyi P M B. The google similarity distance[J]. Knowledge and Data Engineering，IEEE Transactions on，2007，19(3)：370-383.

[12] Wang L，Li S，Wong D F，et al. A joint chinese named entity recognition and disambiguation system[C]//Proceedings of The 2nd CIPSSIGHAN Joint Conference on Chinese Language Processing（CLP-2012）. 2012：146-151.

[13] Liu J，Xu R，Lu Q，et al. Explore chinese encyclopedic knowledge to disambiguate person names[C]//Proceedings of The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing（CLP-2012）. 2012：138-145.

[14] Han W，Liu G，Mao Y，et al. Attribute based Chinese Named Entity Recognition and Disambiguation[C]//Proceedings of The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing（CLP-2012）. 2012：127-131.

[15] Peng Z，Sun L，Han X. SIR-NERD：A Chinese Named Entity Recognition and Disambiguation System using a Two-Stage Method[C]//Proceedings of The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing（CLP-2012）. 2012：114-120.

[16] Minkov E，Cohen W W，Ng A Y. Contextual search and name disambiguation in email using graphs[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM，2006：27-34.

[17] Bekkerman R，McCallum A. Disambiguating web appearances of people in a social network[C]//Proceedings of the 14th international conference on World Wide Web. ACM，2005：463-470.

[18] 郎君，秦兵，宋巍，等. 基于社会网络的人名检索结果重名消解[J]. 计算机学报，2009，32(7)：1365-

1374.

[19] Hoffart J, Seufert S, Nguyen D B, et al. Kore: Keyphrase overlap relatedness for entity disambiguation [C] //Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012: 545-554.

[20] Ikeda M, Ono S, Sato I, et al. Person name disambiguation on the web by two-stage clustering[C]// Proceedings of the 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference. 2009.

[21] E Elmacioglu, Y Tan, S Yan, et al. PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features[C] //Proceedings of The SemEval-2007, 2007: 268-271.

[22] Och F J. Minimum error rate training in statistical machine translation[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003: 160-167.

[23] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[24] **刘挺，车万翔，李正华. 语言技术平台**[J]. **中文信息学报**, 2012, 25(6): 53-62.

[25] http://genes. toronto. edu/index. php ?q = affinity% 20propagation[OL].

[26] Hao Zong, Derek F Wong, Lidia S Chao. A template based hybrid model for chinese personal name disambiguation [C]//Proceedings of The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012). 2012: 121-126.

线岩团(1981—)，**博士研究生，讲师，主要研究领域为自然语言处理、信息抽取。**
E-mail: xianyantuan@qq. com

洪旭东(1989—)，**博士研究生，主要研究领域为自然语言处理、信息检索。**
E-mail: 459102534@qq. com

余正涛(1970—)，**通信作者，博士，教授，博士生导师，主要研究领域为自然语处理、信息检索、机器翻译、机器学习等。**
E-mail: ztyu@hotmail. com