Journal of Computer Applications
计算机应用,2013,33(4):984-987
ISSN 1001-9081
CODEN JYIIDU
2013-04-01
http://www.joca.cn

文章编号:1001-9081(2013)04-0984-04
doi:10.3724/SP.J.1087.2013.00984

基于支持向量机分类和语义信息的中文跨文本指代消解

赵知纬[1,2],顾静航[1,2],胡亚楠[1,2],钱龙华[1,2*],周国栋[1,2]
(1.苏州大学 自然语言处理实验室,江苏 苏州 215006; 2.苏州大学 计算机科学与技术学院,江苏 苏州 215006)
(*通信作者电子邮箱 qianlonghua@suda.edu.cn)

摘 要:跨文本(实体)指代消解(CDCR)的任务就是把所有分布在不同文本但指向相同实体的词组合在一起形成一个指代链。传统的跨文本指代消解主要采用聚类方法来解决信息检索中遇到的重名消歧问题。将聚类问题转换为分类问题,并采用支持向量机(SVM)分类器来解决信息抽取中的重名消歧和多名聚合问题。该方法可有效融合实体名称的构词特征、读音特征以及文本内部和文本外部的多种语义特征。在中文跨文本指代语料库上的实验表明,同聚类方法相比,该方法在提高精度的同时,也提高了召回率。

关键词:跨文本指代;信息抽取;支持向量机分类器;语义信息;重名消歧;多名聚合
中图分类号:TP391 **文献标志码**:A

Chinese cross document co-reference resolution based on SVM classification and semantics

ZHAO Zhiwei[1,2], GU Jinghang[1,2], HU Yanan[1,2], QIAN Longhua[1,2*], ZHOU Guodong[1,2]
(1. *Laboratory of Natual Language Processing, Soochow University, Suzhou Jiangsu* 215006, *China*;
2. *School of Computer Science and Technology, Soochow University, Suzhou Jiangsu* 215006, *China*)

Abstract: The task of Cross-Document Co-reference Resolution (CDCR) aims to merge those words distributed in different texts which refer to the same entity together to form co-reference chains. The traditional research on CDCR addresses name disambiguation posed in information retrieval using clustering methods. This paper transformed CDCR as a classification problem by using an Support Vector Machine (SVM) classifier to resolve both name disambiguation and variant consolidation, both of which were prevalent in information extraction. This method can effectively integrate various features, such as morphological, phonetic, and semantic knowledge collected from the corpus and the Internet. The experiment on a Chinese cross-document co-reference corpus shows the classification method outperforms clustering methods in both precision and recall.

Key words: cross document co-reference resolution; information extraction; Support Vector Machine (SVM) classifier; semantics; name disambiguation; variant consolidation

0 引言

跨文本指代消解(Cross Document Coreference Resolution, CDCR)任务面临两项挑战:重名消歧与多名聚合。前者是将同一名称的不同实体区分开来,即布什既可表示美国第43任总统乔治·W·布什,也可表示美国第41任总统乔治·H·W·布什;而后者是将指向同一实体的不同名称合并起来,如北韩与朝鲜都表示同一个实体——朝鲜民主主义共和国。随着自然语言处理技术从单个文本内的信息抽取向多文本间的信息融合方向发展,作为文本间信息连接的重要纽带,跨文本指代消解引起了广泛的研究兴趣。

传统的跨文本指代消解主要面向信息检索,即将检索到的Web网页中具有相同名称的不同实体区分开来,其本质是重名消歧问题[1-2]。随着信息抽取技术的日渐成熟以及信息融合需求的不断加强,面向信息抽取的跨文本指代消解的研究得到了广泛的重视[3]。与面向信息检索的跨文本指代消解不同的是,后者不仅要解决重名消歧问题,还要解决更为严重的多名聚合问题[4-5]。传统的跨文本指代消解采用基于聚类的无监督方法,其优点是无需训练语料,也取得了较好的性能[1,6-11]。不过,它的缺点在于无法综合考虑影响指代消解的多种因素和特征,因而将聚类算法应用于面向信息抽取的跨文本指代消解时,效果不尽理想[4,12]。

本文首先在ACE2005中文语料库的基础上,通过自动生成和手动标注相结合办法构建了一个面向信息抽取的涵盖所有ACE实体类型的中文CDC语料库。然后,在该语料上,利用机器学习的分类方法充分融合各种构词特征、读音特征和语义特征,从而显著提高跨文本指代消解的性能。

1 问题定义

实体是指现实世界中存在的对象,如人物、机构和组织等,而实体出现在自然语言文本中的名称,则称为实体表述。一般来说,实体与实体表述是以一种多对多的关系,即一个实体(e_j)可能有一个或多个表述(m_j),反过来,一个表述可能对应多个实体。

定义D为一个文档的集合,E^D为该文档集之上的实体集合,M^D为文档集中出现的表述集合,即:

$E^D = \{e_1, e_2, \cdots, e_i, \cdots, e_n\}$,如中国、朝鲜、韩国等。

$M^D = \{m_1, m_2, \cdots, m_j, \cdots, m_b\}$,如北朝鲜、北韩、朝鲜、南

收稿日期:2012-09-24;**修回日期**:2012-11-27。 **基金项目**:国家自然科学基金资助项目(60873150,90920004);江苏省自然科学基金资助项目(BK2010219);江苏省高校自然科学重大项目(11KJA520003)。

作者简介:赵知纬(1987-),男,浙江杭州人,硕士研究生,主要研究方向:信息抽取; 顾静航(1987-),男,河南洛阳人,硕士研究生,主要研究方向:信息抽取; 胡亚楠(1989-),女,安徽亳州人,硕士研究生,主要研究方向:信息抽取; 钱龙华(1966-),男,江苏苏州人,副教授,CCF会员,主要研究方向:自然语言处理; 周国栋(1967-),男,江苏溧阳人,教授,博士生导师,CCF高级会员,主要研究方向:自然语言处理。

朝鲜、南韩、韩国等。

值得注意的是,E^D 中的所有实体都是不同的,而 M^D 中的实体表述则可能是相同的,即同一个表述出现在文档集的不同位置。

跨文本指代消解就是将文档集中属于同一实体的不同表述合并起来,归入到同一个指代链中,即找到一个映射 f,使得 $f(m_j) = e_i(m_j \in M^D, e_i \in E^D)$,且从 M^D 到 E^D 是单射的。实际情况下,对于某一个文档集合,实体集虽然客观存在但对本研究而言是未知的,因此跨文本指代消解的任务实际上就是从实体表述到实体的一个聚类问题。

2 基于分类的跨文本指代消解

传统的跨文本指代消解普遍采用聚类的方法,其方法流程是首先抽取实体表述的上下文信息、句法信息及实体所在文档的元数据等作为实体表述的特征,然后利用这些特征来计算实体表述之间的总体相似度从而构建起一个相似度矩阵,最后使用聚类算法,如层次聚类、模糊聚类等合并相似度较高的簇。这样,一个簇中的实体表述即为一条跨文本指代链。

由于跨文本指代现象涉及的特征众多,它不但与表述所处的上下文信息相关,还与文档中共现的其他实体表述相关,且不同特征所起的作用也不同。例如"巴勒斯坦"和"巴基斯坦"词形高度相似,却非同一实体的不同表述;而"澳大利亚"和"澳洲"虽词形差异较大,但经常会跟"墨尔本"、"悉尼"一同出现。在聚类过程中计算实体表述相似度时难以合理调节各个特征相似度之间的权重,另外对于那些二元特征,如公共字符是否以相同顺序出现,聚类方法也不能很好地合并到总体相似度中。而统计机器学习方法,如支持向量机(Support Vector Machine,SVM)分类器[13],可以从训练语料库中自动训练得到一个模型,该模型能有效融合不同的特征,且考虑不同特征的权重差异。鉴于此,本文提出了用分类的方法来实现跨文本指代消解。

2.1 基本思路

对于面向信息抽取的跨文本指代消解任务,初步实验表明,采用简单的字符串精确匹配方法可以达到 96.8% 的精度,而其召回率只有 82.4%。这说明实体重名歧义现象较少,而多名指代现象相对较严重。因此,本文首先使用精确匹配方法来解决重名歧义问题,然后在该基础上再解决多名指代问题。

由于分类器所处理的是二元或多元分类问题,如判断两个实体表述是否指向同一个实体,而跨文本指代消解的最终目的是要形成一条条包含实体表述的指代链,一条指代链中的所有实体表述都指向同一实体。因此,可通过下列过程将聚类问题转化为二元分类问题:

1)对于所有待消解的实体表述,如 $\{m_1, m_2, \cdots, m_j, \cdots, m_k\}$,采用字符串精确匹配方法合并,得到新的实体表述集合 $\{m'_1, m'_2, \cdots, m'_j, \cdots, m'_l\}$;

2)将这些实体表述两两配对形成不重复的二元组:

$\{(m'_i, m'_j)\}_{i,j=1}^{l}; i < j$

这样指代链的构建就转变成判断这些二元组中的每一对实体表述是否指向同一实体的分类问题。

3)将这些二元组由训练好的分类器模型来判断是否指向同一实体,最后利用传递关系将指向同一实体的表述链接到一起形成指代链。例如,假设有下列三对表述 $\{(m_1, m_2), (m_1, m_3), (m_2, m_3)\}$ 指向同一实体,最后形成的指代链为 $\{(m_1, m_2, m_3)\}$。

2.2 特征选择

除了选择恰当的分类器之外,决定分类性能的关键因素之一是分类特征项的选择。对于两个表述是否属于同一实体的二元分类问题,本文选取了构词特征、读音特征和语义特征等三大特征集,每类特征集分成多个特征。表 1 列出了这些特征的类型、特征名称、特征描述和特征值的类型。

需要说明的是:

1)构词特征之 Ch-Modifier-Match:词后缀是指如 FAC、GPE、LOC、ORG、VEH、WEA 等实体名称的中心词,它通常表明一定的属性,如"雅虎公司"中的"公司"等。首先将所有实体表述进行分词,提取其分词结果的最右部分作为初始后缀。如果该后缀出现的频率大于某一阈值,则将它加入后缀列表。

2)读音特征:由于拼音输入法的广泛使用,某些实体表述中含有错拼的汉字。本文使用 Excel 宏来获取实体表述所对应的汉字拼音。

3)外部语义资源之 Google-Spell-Correct:语料库中存在由拼写错误而导致的多名现象。在查询实体表述时,Google 搜索引擎会根据其查询历史、查询词之间的相关度等提示可能存在的拼写错误。

4)外部语义资源之 Wiki-Anchor-Text:维基百科里的锚点文本与条目之间的链接关系蕴含了丰富的知识。同一实体表述 m 可以被链接到不同的条目 t,这样一个表述可表示为由其条目所构成的向量,即:

$$m = \{(t_0, c_0), (t_1, c_1), \cdots, (t_i, c_i), \cdots, (t_n, c_n)\}$$

其中:t_i 表示被链接到的条目,c_j 表示被链接至条目 t_i 的次数。于是两个表述的相似度为:

$$sim(m_1, m_2) = \sum_{t_{ik}=t_{jk}} \frac{c_{ik}}{\sqrt{\sum c_i^2}} \times \frac{c_{jk}}{\sqrt{\sum c_j^2}}$$

5)外部语义资源之 Tyccl-Code:《同义词词林》中的每个词语都含有一个 8 位编码,以映射其语义类别。本文采用了前 5 位的编码,以避免过细的语义颗粒度。

3 实验

3.1 语料库

本文采用的语料库是在 ACE2005 中文语料的基础上构建的跨文本指代语料库。该语料库共有 633 篇文章,分别选自新华网、央视、台北国际之声、联合早报、马来西亚之声等境内外中文媒体的新闻、博客、演讲、访谈等。语料库中总共有 3618个实体和 6771 个实体表述。

在语料库的 633 个文件中随机挑选 317 个文件组成训练集,剩下的文件构成测试集。其中包含了 3442 个实体表述,2080个实体。这些实体表述中,GPE 类型占 1239 个,PER 类型占 967 个,FAC 类型占 146 个,LOC 类型占 117 个。

3.2 语料库过滤

基于分类的跨文本指代消解需要将全部的实体表述两两配对,导致形成的二元组的数量高达百万级。但事实上,其中有相当大比例的二元组是没有指代关系的,全部保留这些二元组将会极大地影响分类器的分类性能。因此本文制定了若干过滤规则,来去除那些不太可能有指代关系的二元组,即满足下列规则的二元组才保留下来:

1)两个表述必须属于相同的实体类型;

2)两个表述所对应的字符串的字符 Dice 系数必须大于某一给定阈值;

3)如果两个表述在字面上完全不同,那么它们必须符合从互联网挖掘到的可能的别名关系。

在生成初始训练实例和测试实例后,通过实施上述过滤策略后得到最终的训练实例和测试实例,其中训练语料中正例和负例的比例约为1:10.6,大大缓解了训练语料中的类别不平衡现象;同时测试实例最终数目为2336对,因而也极大地减少了测试实例。由于所需要的是二元分类器,因此选择 SVM^{Light}[16]分类器。

表1 跨文本指代消解的特征集合

特征类型		特征名称	特征描述	特征值类型
构词特征		Ch-Unigram	一元字符的 Dice 系数	实数
		Ch-Bigram	二元字符的 Dice 系数	实数
		Ch-Order	词对间的公共字符是否同序出现	布尔
		Ch-Modifier-Match	去掉词后缀剩余修饰部分是否完全匹配	布尔
		Ch-Part-Match	一个词是否与另一个词的词头或者词尾完全匹配	布尔
		Ch-Substring	一个字符串是否为另一个字符串的子集	布尔
读音特征		Pron-Trigram	三元字符的 Dice 系数	实数
		Pron-Fourgram	四元字符的 Dice 系数	实数
		Pron-Fivegram	五元字符的 Dice 系数	实数
		Pron-Exact-Match	两个词对应的拼音是否完全相同	布尔
		Pron-Order	字符串间的所有公共子串是否同序出现	布尔
		Pron-Modifier-Match	去掉词后缀剩余修饰部分的拼音是否完全相同	布尔
语义特征	文本内部	Within-Doc-Coref	两个词是否在某一文本内存在文本内指代关系	布尔
		Ent-Type	实体的类型(如 PER、ORG 和 GPE 等)	枚举
		Co-occur-Ent-Num	相同的共现实体数量是否大于2	布尔
	外部资源	Google-Co-occurr-Ent	共现实体通过搜索引擎计算得到的 Jaccard 系数[14]	实数
		Google-Spell-Correct	搜索引擎是否提供有拼写矫正	布尔
		Wiki-Anchor-Text	根据维基百科内锚点文本和条目的超链接计算得到的词的相似度[15]	实数
		Wiki-Redirection	是否在维基百科中存在页面重定向	布尔
		Tyccl-Code	词后缀在同义词词林中是否有相同的编码	布尔

3.3 实验结果和分析

本文采用 B3 打分标准[17]来计算精度(P)和召回率(R),再通过 $F1 = 2 \cdot P \cdot R/(P+R)$ 来评价系统的总体性能。

3.3.1 不同特征的影响

表2比较了不同类型的特征对于跨文本指代消解(CDCR)性能的影响,其中各类特征是以累加的方式加入特征集的。从表中可以看出:

表2 不同特征对 CDCR 的性能影响 %

特征集	精度	召回率	$F1$ 指数
构词特征	93.6	87.5	90.5
+读音特征	94.8	85.3	89.8
+文本内部语义(文本内指代、实体类型、共现实体数)	94.2	89.2	91.6
+Google(共现实体 Jaccard 系数、拼写矫正)	94.3	89.2	91.7
+维基百科(条目重定向、超链接相似度)	94.6	92.8	93.7
+同义词词林	95.0	93.1	94.0

1)在加入各种类型的特征后,跨文本指代消解的性能获得了显著的提高,$F1$ 指数达到了94.0%,特别是召回率获得了极大的提高(5.6%),这说明这些特征对跨文本指代消解具有一定的作用。

2)仅使用构词特征时,精度较高而召回率相对低一点,这进一步印证了语料库中重名现象较少而多名现象较严重的特点。

3)读音特征虽然降低了系统的召回率,但略微提高了精度。其原因在于读音特征能将"首都国际机场"和"高雄国际机场"等错误合并的词对分开的同时,也会将如"北京"和"北京市"等正确合并的词对分开。但是在语义信息加入之后,读音特征会对整体性能有促进作用,因此仍保留该特征。

4)语义特征对整个系统的贡献度是最高的,在保证精度不损失的情况下,显著提高了召回率。因为文本内部和外部语义特征,如文本内部的指代关系以及维基百科中的重定向信息等,对两个表述是否指向同一实体具有重要的作用。

3.3.2 与聚类方法的比较

表3比较了采用精确匹配、聚类方法和分类方法进行跨文本指代消解的性能。由于聚类方法的局限性,它无法应用表1的所有特征。为了尽可能公平并提高性能,采用单连通的层次聚类算法,其总体相似度的计算方法如下:

1)相似度为表1中的所有实数特征的算术平均;

2)当两个实体表述在维基百科中存在重定向页面或者文本内相互指代,则将相似度置为1;

3)两实体表述的类型若不相同,则相似度置为0;

4)两实体表述的字符 Dice 系数若小于给定阈值,则相似度置为0。

表3的实验结果表明:

1)同精确匹配方法和聚类方法相比,分类方法在跨文本指代消解任务中取得了最好的性能,其 $F1$ 指数分别比两者高出5%和2.8%,且精度损失较小甚至还有提高。这是由于基于统计学习的分类器能有效融合各种相似度特征和二元特征,并且能自动调整不同特征的权值以反映不同特征的贡献度,而聚类方法难于有效调整各种相似度之间的权值。

2)同精确匹配方法相比,聚类方法也能在一定程度上提高 $F1$ 指数,不过,在提高召回率的同时,精度也在明显下降。对于跨文本指代消解来说,精度下降意味着不同实体的表述

被错误地合并在一起,从而损害了后期跨文本信息融合的准确性。从这一点上来说,分类方法也优于聚类方法。

表 3　不同方法的 CDCR 性能比较　%

方法	精度	召回率	F1 指数
精确匹配	96.8	82.4	89.0
聚类方法	93.7	88.8	91.2
分类方法	95.0	93.1	94.0

4　结语

本文采用 SVM 分类器的方法,通过融合实体表述的各种构词特征、读音特征和语义特征,来解决跨文本指代消解问题。在跨文本指代消解语料库上的实验表明,分类方法优于传统的聚类方法,且文本内部和文本外部的语义信息可以显著提高跨文本指代消解的性能。一方面,相比聚类方法,分类方法可以考虑更多难于结合到相似度中的离散化特征,还可以灵活调整各类特征之间的权重;另一方面,语义特征可以使跨文本指代消解在保持较高精度的前提下,进一步提高召回率。

今后的工作,一方面将尝试在大规模语料库上进行跨文本指代消解的实验;另一方面将采用更有效的特征来提高系统的精度,因为精度在信息融合中是至关重要的。

参考文献:

[1] MCCARTHY L W. Using decision trees for coreference resolution [C]// MUC-6: Proceedings of the Sixth Message Understanding Conference. Montreal, Quebec, Canada: [s. n.], 1995: 20 - 25.

[2] BAGGA A, BALDWIN B. Entity-based cross-document coreferencing using the vector space model [C]// COLING-ACL'98: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 1998: 79 - 85.

[3] NIST speech group. The ACE2008 evaluation plan: assessment of detection and recognition of entities and relations within and across documents [EB/OL]. [2008 - 08 - 08]. http://www. nist. gov/speech/tests/ace/2008/doc/ace08-evalplan. v1. 2d. pdf.

[4] BARON A, FREEDMAN M. Who is who and what is what: experiments in cross-document co-reference [C]// EMNLP'08: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008: 274 - 283.

[5] SINGH S, SUBRAMANYA A, PEREIRA F, *et al*. Large-scale cross-document coreference using distributed inference and hierarchical models [C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011: 793 - 803.

[6] GOOI C H, ALLAN J. Cross-document coreference on a large scale corpus [C]// HLT-NAACL 2004. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004: 9 - 16.

[7] BOLLEGALA D, MATSUO Y, ISHIZUKA M. Disambiguating personal names on the Web using automatically extracted key phrases [C]// Proceedings of the European Community of Artificial Intelligence. [S. l.]: IOS Press, 2006: 553 - 557.

[8] HUANG JIAN, TAYLOR S M, SMITH J L, *et al*. Profile based cross-document coreference using kernelized fuzzy relational clustering [C]// Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009: 414 - 422.

[9] POPESCU O. Person cross document coreference with name perplexity estimates[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009: 997 - 1006.

[10] POPESCU O. Dynamic parameters for cross document coreference [C]// COLIN 2010. Beijing: [s. n.], 2010: 988 - 996.

[11] CHEN Y, JIN P, LI W J, *et al*. The Chinese persons name disambiguation evaluation: exploration of personal name disambiguation in Chinese news [C/OL]// Joint Conference on Chinese Language Processing 2010. Beijing: ACL, 2010[2012 - 09 - 01]. ttps://www. aclweb. org/anthology-new/W/W10/W10-4152. pdf.

[12] LLOYD L, MEHLER A, SKIENA S. Identifying co-referential names across large corpora [C]// Combinatorial Pattern Matching. Barcelona, Spain: [s. n.], 2006: 12 - 23.

[13] JOACHIMS T. Making large-scale SVM learning practical [J]. Advances in Kernel Methods - Support Vector Learning. Cambridge: MIT Press, 1999.

[14] KALASHNIKOV D V, NURAY-TURAN R, MEHROTRA S. Towards breaking the quality curse. a Web-querying approach to Web people search [C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2008: 27 - 34.

[15] HAN XIANPEI, ZHAO JUN. Named entity disambiguation by leveraging Wikipedia semantic knowledge [C]// Proceeding of the 18th ACM Conference on Information and Knowledge Management. New York: ACM 2009: 215 - 224.

[16] JOACHIMS T. Thorsten Joachims' Home Page [EB/OL]. [2010 - 05 - 05]. http://svmlight. joachims. org/.

[17] BAGGA A. Evaluation of coreferences and coreference resolution systems [C]// Proceedings of the First Language Resource and Evaluation Conference. Granada, Spain: [s. n.], 1998: 563 - 566.

(上接第 952 页)

[2] 熊志强,黄佳庆,刘威,等. 无线网络编码综述[J]. 计算机科学,2007,34(13):6 - 9.

[3] FRAGOULI C, BOUDEC J Y L, WIDMER J. Network coding: an instant primer[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(1): 63 - 68.

[4] AHLSWEDER, CAI N, LI S R, *et al*. Network information flow [J]. IEEE Transactoins on Information Theory, 2000, 46(4): 1204 - 1216.

[5] LI S Y R, YEUNG R W, CAI N. Linear network coding[J]. IEEE Transactions on Information Theory, 2003, 49(2): 371 - 381.

[6] KOETTER R, MEDARD M. An algebraoc approach network coding[J]. IEEE/ACM Transactions on Networking, 2003, 11(5): 782 - 795.

[7] CHOU P A, WU Y, JAIN K. Practical network coding [C]// Proceedings of the Annual Allerton Conference on Communication, Control and Computing. Monticello: [s. n.], 2003: 473 - 482.

[8] JAGGI S, SANDERS P, CHOU A, *et al*. Polynomial time algorithms for multicast network code construction[J]. IEEE Transactions on Information Theroy, 2005, 51(6): 1973 - 1982.

[9] HO T, MEDARD M, KOETTER R, *et al*. A random linear network coding approach to multicast[J]. IEEE Transactions on Information Theory, 2006, 52(10): 4413 - 4430.

[10] 蒲保兴,王伟平. 线性网络编码运算代价的估算与分析[J]. 通信学报,2011,32(5):47 - 55.

[11] WANG M, LI B C. How practical is network coding?[C]// IWQos 2006: The 14th IEEE International Workshop on Quality of Service. Piscataway, NJ: IEEE Press, 2006: 274 - 278.

[12] 王兵山. 离散数学[M]. 长沙:国防科技大学出版社,2004:263 - 281.

[13] 谢金星,邢文训,王振波. 网络优化[M]. 北京:清华大学出版社,2009:79 - 82.

基于支持向量机分类和语义信息的中文跨文本指代消解

作者： 赵知纬， 顾静航， 胡亚楠， 钱龙华， 周国栋， ZHAO Zhiwei， GU Jinghang， HU Yanan， QIAN Longhua， ZHOU Guodong

作者单位： 苏州大学自然语言处理实验室,江苏苏州215006;苏州大学计算机科学与技术学院,江苏苏州215006

刊名： 计算机应用 ISTIC PKU

英文刊名： Journal of Computer Applications

年，卷(期)： 2013,33(4)

参考文献(17条)

1. MCCARTHY L W Using decision trees for coreference resolution 1995

2. BAGGA A;BALDWIN B Entity-based cross-document coreferencing using the vector space model 1998

3. NIST speech group The ACE2008 evaluation plan:assessment of detection and recognition of entities and relations within and across documents 2008

4. BARON A;FREEDMAN M Who is who and what is what:experiments in cross-document co-reference 2008

5. SINGH S;SUBRAMANYA A;PEREIRA F Large-scale cross-document coreference using distributed inference and hierarchical models 2011

6. GOOI C H;ALLAN J Cross-document coreference on a large scale corpus 2004

7. BOLLEGALA D;MATSUO Y;ISHIZUKA M Disambiguating personal names on the Web using automatically extracted key phrases 2006

8. HUANGJIAN;TAYLOR S M;SMITH J L Profile based cross-document coreference using kernelized fuzzy relational clustering 2009

9. POPESCU O Person cross document coreference with name perplexity estimates 2009

10. POPESCU O Dynamic parameters for cross document coreference 2010

11. CHEN Y;JIN P;LI W J The Chinese persons name disambiguation evaluation:exploration of personal name clisambiguation in Chinese news 2010

12. LLOYD L;MEHLER A;SKIENA S Identifying co-referential names across large corpora 2006

13. JOACHIMS T Making large-scale SVM learning practical 1999

14. KALASHNIKOV D V;NURAY-TURAN R;MEHROTRA S Towards breaking the quality curse.a Web-querying approach to Web people search 2008

15. HAN XIANPEI;ZHAO JUN Named entity disambiguation by leveraging Wikipedia semantic knowledge 2009

16. JOACHIMS T Thorsten Joachims' Home Page 2010

17. BAGGA A Evaluation of coreferences and coreference resolution systems 1998