第37卷 第18期 计 算 机 工 程 2011年9月
Vol.37 No.18 Computer Engineering September 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)18—0180—03 文献标识码：A 中图分类号：TP391

基于特征分选策略的中文共指消解方法

李渝勤[1,2]，甘润生[1]，杨永红[3]，施水才[1,2]

(1. 北京信息科技大学计算机学院，北京 100101；2. 北京拓尔思信息技术股份有限公司，北京 100101；
3. 中山大学信息科学与技术学院计算机科学系，广州 510275)

摘 要：针对基于机器学习的中文共指消解中不同类别名词短语特征向量的使用差异，提出一种基于特征分选策略的方法。该方法在选择特征向量时对人称代词和普通名词短语分别处理，充分利用不同名词短语的已有特征进行共指消解，并减少部分无效特征在共指消解过程中产生的“噪声”。实验结果表明，该中文共指消解方法能提高共指消解的性能，*F* 值达到80.72%。

关键词：共指消解；特征选择；自然语言处理；支撑向量机；数据词典

Chinese Coreference Resolution Method Based on Feature Respective Selection Strategy

LI Yu-qin[1,2], GAN Run-sheng[1], YANG Yong-hong[3], SHI Shui-cai[1,2]

(1. Computer School, Beijing Information Science & Technology University, Beijing 100101, China;
2. Beijing TRS Information Technology Co. Ltd., Beijing 100101, China;
3. School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510275, China)

【Abstract】This paper studies different features based up on the type of noun phrase in Chinese coreference resolution based on machine learning, and proposes features selection strategy to be applied to coreference resolution, the approach selects pronouns and other noun phrases features respectively, so this method can reduce some “noise” and utilize features effectively. Experimental results show that the method can improve the performance of coreference resolution system, and *F*-measure reaches 80.72%.

【Key words】coreference resolution; feature selection; nature language processing; Support Vector Machine(SVM); data dictionary

DOI: 10.3969/j.issn.1000-3428.2011.18.059

1 概述

共指现象广泛存在于自然语言的各种表达中，表示篇章中的一个语言单位与之前出现的语言单位存在语义上的关联(本文不讨论回指和零指)，用于指向的语言单位称为照应语，被指向的语言单位为先行语。确定照应语和先行语之间关系的过程就是共指消解，类型一般有人称代词消解和名词短语消解。

共指消解方法早期侧重于理论的研究，在自然语言处理技术发展的带动下，多数研究偏向基于机器学习的方法[1-2]。当前很多的研究更侧重于挖掘更多的有效信息作为特征用于共指消解，研究其对共指消解性能的作用[3]。

相对于英语来说，中文共指消解的研究起步较晚，且还不够深入，但几十年来，学者在相关方面也取得了一定的成果[4-5]。在前人共指消解研究中发现，特征向量的选择对系统的最终性能有很大的影响。使用相同的机器学习方法，但选取不同的特征向量，系统的性能可能有较大的差异。一些研究者在共指消解研究中没有对名词短语的类别进行细分和区别处理，也没有深入考虑人称代词和其他名词短语的差异，在特征选取和处理方式上不尽相同。本文针对人称代词和其他名词短语分别选取特征，对不同特征构建的实例利用支撑向量机(Support Vector Machine, SVM)分类器进行分类。

2 共指消解平台

本文的系统平台是在文献[1]的基础上实现，根据流程构建出如图1所示的共指消解基本框架。从图1可看出，本文的共指消解系统由预处理、特征向量选择、训练和测试实例构造、机器学习算法(SVM)，以及消解结果构成。

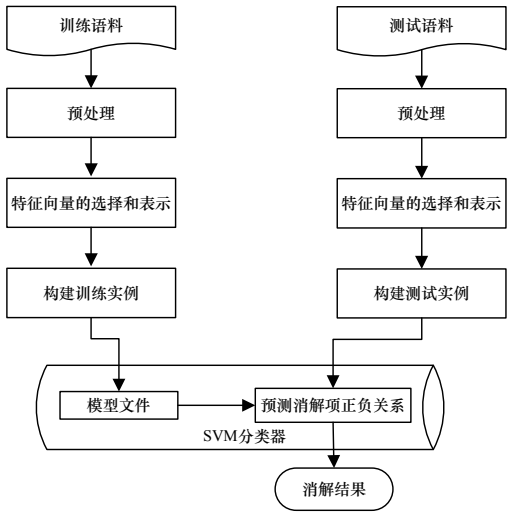

图1 共指消解流程

基金项目：国家“863”计划基金资助重点项目(2006AA010105)；国家自然科学基金资助项目(60772081)；北京市自然科学基金资助项目(4092015)；北京市教委科技发展计划基金资助项目(KM201010772023)

作者简介：李渝勤(1963－)，女，副研究员，主研方向：中文信息处理，信息检索；甘润生，硕士研究生；杨永红，讲师；施水才，教授

收稿日期：2011-03-02 **E-mail：**wsgrs@126.com

为了更好地进行共指消解，本文对文本预处理的解决充分利用了语料中标注的实体链信息。在预处理过程中对源文档只进行分句和获得位置信息，其他信息使用实体链中已标注的内容，预处理中对这些已标注信息进行抽取并组织。

在构建训练和测试集合时进行过滤，这是考虑到一篇文档包括多个名词短语，人称代词有不带性别概念和具体职称的习惯用法，如“玩他个痛快”、“管他三七二十一”等，有时虽然有指代的意思，但篇章中没有具体的指代内容，如“我们再来看看巴以局势”等。人称代词的先行语可以是人、地方、机构的专有名词和指人的普通名词等，语言单位越小，先行语候选就越多。有些名词短语在语篇中只出现一次或出现多次，但所指的实体类别不同。上述情况都需要进行过滤处理，如果简单地将各名词短语两两组对就会生成大量的实例，而且大多名词短语不具有共指关系，这就会产生大量的噪声，影响共指消解的性能和系统处理速度，因此，在生成训练和测试实例之前，首先需要对篇章中的无关名词短语进行过滤，对经过过滤后的先行语集合再生成训练和测试实例，能提高共指消解系统的性能。

3 特征向量

虽然使用更多的特征可能有助于共指消解，但无用特征可能会对一些名词短语有消极影响，产生“噪声”降低共指消解系统的性能。一般在选择人称代词特征时，不需要加入字符串是否匹配或是否为其子串等特征，也无需引入语义相似等无用特征；选取名词短语特征时一般无需判断是否指人，只需看语义类别是否一致即可。本文在选择特征向量时对人称代词和名词短语分别处理。

3.1 人称代词特征选择和表示

中文人称代词数目有限且固定，相对较容易从文档中获得，根据语料中出现的人称代词，本文处理出现在语料中的20个人称代词，包括我(们)、双方、彼此、其等。

汉语中的人称代词“它”和“它们”用来指代非人类相关的实体，其他代词在一般情况都是用来指人：第1人称和第2人称代词基本指代人类实体；第3人称代词“他”和“他们”一般指代性别为男性的人类实体，其中前者表示单数实体，后者表示复数实体，但在实际应用时，“他”和“他们”也用来指代非人实体，而且“他们”也可以指代没有明确复数特征的词语。

本文建立了一个人称代词数据词典，词典包含部分词语、词语对应的能够明显分辨的性别和单复数信息。

人称代词的特征选择如下：

(1)指人匹配：人称代词与候选先行语都指人时，即它们的值都为1时，指人匹配特征值为1，否则特征值为0。

(2)性别匹配：人称代词与候选先行语的性别是否一致，如果其中之一的值为0，性别匹配特征值为0.5，当且仅当它们的值相同，性别匹配特征值为1，否则特征值为0。

(3)单复数匹配：判断人称代词与先行语的单复数是否一致，此特征是进行候选集过滤的有效手段，2个词语的单复数一致，即取值相同，单复数匹配特征值为1，不同特征值为0，其中之一为0，特征值为0.5。

(4)句子距离：一般认为人称代词与先行语的距离近，存在共指关系的可能性更大。照应语与先行语位于同一句，特征值为1；2句特征值为0.8，3句特征值为0.6，4句特征值为0.2，超过4句特征值为0。

(5)位置：本文用的位置信息是照应语与先行语之间的字符距离。与句子距离类似，距离越远，存在共指关系的可能性就越小。通过分析语料，有共指关系的2个名词短语之间的字符距离绝大多数在1个~180个字符之间。照应语与先行语之间的距离在50个字符以内，特征值为1；超过200个字符，特征值为0；在50个和200个字符之间，特征值递减0.2。

(6)实体类别：当人称代词指代的实体类别和先行语的类别一致时，特征值为1，否则特征值为0。

3.2 其他名词短语特征选择和表示

名词短语相对复杂，存在多种关系。进行有效消解需要更多的特征，本文选择的特征除了上述之外，还包括如下特征：

(1)字符串匹配：照应语与先行语的字符完全相同，即完全匹配，特征值为1，如果两词之间存在包含关系，即其中之一为子串，特征值为0.5，其他情况特征值为0。

(2)中心词匹配：通过式(1)计算照应语与先行语的中心词匹配程度，特征值为从0~1的封闭集实数。计算方式如下：

$$match(n_1,n_2)=\frac{2\times same(n_1,n_2)}{len(n_1)+len(n_2)} \tag{1}$$

其中，$same(n_1,n_2)$判断2个中心词的字符相同的数目；$len(n_1)$表示中心词的字符长度。

(3)别名：为了更有效地应用此特征，本文针对语料内容建立一个别名数据词典，如“巴勒斯坦”与“巴”存在别名关系。照应语与先行语是别名关系，特征值为1，否则特征值为0。

(4)实体提及类别：实体提及的类别有NAM、NOM和PRO 3类。照应语与先行语是相同的类别，特征值为1，否则特征值为0。

4 实验结果与分析

本文使用的语料是ACE2007 broadcast news，共521篇，其中416篇作为训练语料，其余105篇作为测试语料。如表1所示。

表1 实验使用的ACE broadcast news语料

名称	新闻广播语料	训练	测试
文档数量	521	416	105
句子数量	4 526	3 664	862
名词短语数量	22 484	18 042	4 422

使用的SVM分类器是SVM[light1]，参数是默认值。对共指消解系统性能的评价采用准确率、召回率和F值。准确率P是共指消解结果中正确消解的对象数目占实际消解的对象数目的百分比，反映了共指消解系统的准确程度；召回率R是共指消解结果中正确消解的对象数目占消解系统应消解对象总数的百分比，反映了共指消解系统的完备性；F值是对两者的综合考虑。定义如下：

$$F=\frac{(\beta+1)P\times R}{\beta\times P+R}\quad \beta=1 \tag{2}$$

表2给出了共指消解平台在语料上的实验结果，特征向量没有针对人称代词和名词短语区分。SVM有多种核函数，且使用不同的函数得到的结果也会有所不同。

表2 不同核函数未分别选取特征的实验结果 (%)

核函数	准确率	召回率	F值
线性核函数	82.53	65.28	72.09
多项式核函数	85.65	68.94	76.39
径向基核函数	83.77	75.98	79.68

从表2的实验结果可以看出，本文构建基准平台的准确

率、召回率、F 值与与文献[1](58.6%, 67.3%, 62.6%)和文献[2](64.1%, 74.9%, 69.1%)相比，性能有较大幅度的提升，一方面是由于使用的分类器和语料不同，另一方面很大程度上是由于预处理的过程中没有进行分词和词性标注，也没有名词短语识别的丢失和错误，构建训练和测试的部分内容使用的都是语料的实体链中已标注的信息，相对更为精确。从表 2 可以看出，使用径向基核函数与其他 2 个相比准确率有所下降，但召回率有大幅提升，F 值与线性核函数相比增加了超过 7 个百分点。在一系列实验中发现使用径向基核函数得到的效果最好，因此本文使用径向基核函数得到的结果。

人称代词在语料中的数量较少，经统计，语料中的人称代词共有 1 888 个，占所有名词短语的 8%，在训练和测试实例中都只有很少一部分，因此本文在实验中没有对人称代词细分。表 3 是针对人称代词和其余名词短语分别选取特征后的实验结果。表 2 和表 3 的对比如图 2 所示。

表 3 分别选取特征的实验结果 (%)

核函数	准确率	召回率	F 值
径向基核函数	83.89	77.79	80.72

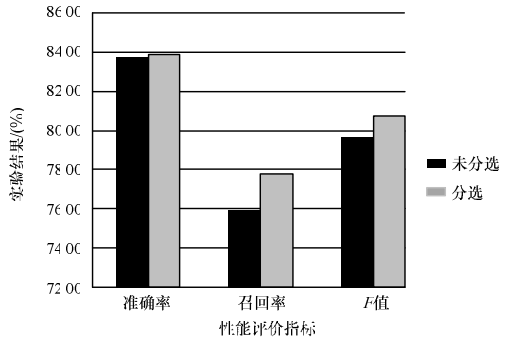

图 2 基于 SVM 的分类结果对比

从图 2 中可以明显看出，与基准系统相比，采用特征分选策略后 F 值和召回率都有所提高，准确率几乎持平，其中召回率提高了近 2%，F 值提高了 1.04%，达到 80.72%。说明对人称代词和其余名词短语分别选取特征向量，能够使系统识别出更多的共指关系，同时也有更多正确的共指关系被识别，从而实现了召回率的提高，证明特征分选策略对共指消解有帮助作用。虽然人称代词较少，在训练和测试中占了很少的比例，但 F 值提高了 1 个百分点，从一定的角度也说明了特征对共指消解的重要影响。

5 结束语

本文通过分析人称代词和其他名词短语所需特征向量的差异性，提出了特征分选策略，实验表明该方法有较好的分类效果，提高了共指消解系统的效率和性能。说明特征分选策略对共指消解性能有增强作用，同时也体现了特征在机器学习方法中的重要作用。对语料中的不同类型名词短语加入不同特征使召回率和 F 值有明显的提高。

但是在当前实验使用的特征仍不够充分，没有进行深入挖掘，如进行句法分析、加入语义信息(语义关系、语义相似度等)等有效特征。下一步工作将考虑获得语义信息，并将其表示作为特征加入共指消解系统，研究其对共指消解的影响。

参考文献

[1] Ng S H, Lim D. A Machine Learning Approach to Coreference Resolution of Noun Phrases[J]. Computational Linguistics, 2001, 27(4): 521-544

[2] Cardie N C. Improving Machine Learning Approaches to Coreference Resolution[C]//Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA: [s. n.], 2002: 104-111.

[3] Ng V. Semantic Class Induction and Coreference Resolution[C]// Proc. of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech: [s. n.], 2007: 536-543.

[4] 王厚峰, 何婷婷. 汉语中人称代词消解研究[J]. 计算机学报, 2001, 24(2): 136-143.

[5] 李艳翠, 杨 勇, 周国栋, 等. 基于支持向量机的英语名词短语指代消解[J]. 计算机工程, 2009, 35(3): 199-201.

编辑 任吉慧

(上接第 179 页)

参考文献

[1] Baker B M, Ayechew M A. A Genetic Algorithm for the Vehicle Routing Problem[J]. Computers & Operations Research, 2003, 30(5): 787-800.

[2] Gillitt B E, Miller L R. A Heuristic Algorithm for the Vehicle Dispatch Problem[J]. Operations Research, 1974, 22(2): 340-349.

[3] Filipec M, Skrlec D, Krajcar S. An Efficient Implementation of Genetic Algorithms for Constrained Vehicle Routing Problem[C]// Proc. of IEEE International Conference on Systems, Man, and Cybernetics. San Diego, USA: [s. n.], 1998: 2231-2236.

[4] Gupta A, Krishnamurti R. Parallel Algorithms for Vehicle Routing Problems[C]//Proc. of the 4th International Conference on High-performance Computing. Amsterdam, Holand: [s. n.], 1997: 144-151.

[5] Jih W R, Hsu Y J. Dynamic Vehicle Routing Using Hybrid Genetic Algorithms[C]//Proc. of IEEE International Conference on Robotics and Automation. Piscataway, USA: [s. n.], 1999: 453-458.

[6] 李 军, 郭耀煌. 物流配送车辆优化调度理论与方法[M]. 北京: 中国物资出版社, 2001.

[7] 张启义. 军事物流运输成本模型研究及应用[D]. 南京: 解放军理工大学, 2008.

[8] 杨 弋, 顾幸生. 物流配送车辆优化调度的综述[J]. 东南大学学报: 自然科学版, 2003, 33(增刊): 105-111.

[9] 姜大力, 杨西龙, 杜 文. 车辆路径问题的遗传算法研究[J]. 系统工程理论与实践, 1999, 19(6): 40-44.

[10] 崔明义. 基于小波消噪变异的浮点数编码遗传算法[J]. 计算机工程, 2010, 36(2): 24-26.

[11] 陈 杰. 基于遗传算法的车辆调度问题解决方案[D]. 天津: 天津大学, 2005.

编辑 任吉慧