第 35 卷 第 3 期 计 算 机 工 程 2009 年 2 月
Vol.35 No.3 Computer Engineering February 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)03—0199—03 文献标识码：A 中图分类号：TP18

基于支持向量机的英语名词短语指代消解

李艳翠，杨 勇，周国栋，朱巧明
(1. 苏州大学计算机科学与技术学院，苏州 215006；2. 江苏省计算机信息处理技术重点实验室，苏州 215006)

摘 要：提出一种基于支持向量机(SVM)的英语名词短语的指代消解方法，并给出具体实现系统。实验采用了几个常用的基本特征，在 MUC-6 公开语料上测试得到的 *F* 值为 68.6，优于同类型的其他原型系统。分析 SVM 中不同核函数对分类结果的影响以及不同的特征对指代消解的作用。实验结果表明，同位语、别名和字符串匹配 3 个特征对指代消解非常重要，距离作为特征使用时对指代消解没有帮助，但可在训练样例生成时作为限制条件来使用。

关键词：指代消解；支持向量机；核函数

Anaphora Resolution of Noun Phrase Based on SVM

LI Yan-cui, YANG Yong, ZHOU Guo-dong, ZHU Qiao-ming
(1. School of Computer Science and Technology, Soochow University, Suzhou 215006;
2. Jiangsu Provincial Key Lab for Computer Information Processing Technology, Suzhou 215006)

【Abstract】This paper proposes an anaphora resolution of noun phrases based on Support Vector Machine(SVM). Evaluation on the MUC-6 corpus using several widely used features shows that the system achieves the *F*-measure of 68.6% and outperforms other similar systems. Further analysis shows that appositive, name alias and full string matching contributes most for anaphora resolution. It also shows that the distance between the antecedent candidate and the anaphor is very useful in constraining the instance generation, although including it as a feature does not help for anaphora resolution.

【Key words】anaphora resolution; Support Vector Machine(SVM); kernel function

1 概述

指代是指在语篇中用一个指代词回指某个以前说到过的语言单位。在语言学中，指代词称为照应语，所指的对象或内容称为先行语，先行语如果在照应语的前面，它们之间的关系称为照应关系称为照应，如果在后面则称为逆照应。指代消解就是确定照应语与先行语之间的相互关系，从而明确照应语指代的是什么对象。本文所研究的指代消解问题只针对照应的情况。

指代大量出现在篇章或对话中，能使句子更加简洁明了，主题更加鲜明突出，但也给计算机理解自然语言增加了难度。在文本摘要、机器翻译、多语言信息处理和信息抽取等诸多应用中都涉及到指代消解问题。1997 年的 EACL 和 1999 年的 ACL 年会设立了指代消解的专题会议[1]，指代消解也是 MUC 和 ACE 信息抽取评测体系中的一个主要任务。指代消解早期的方法侧重于从理论上探索，运用大量手工构建的语言甚至领域知识进行研究。目前多数的指代消解研究趋向于基于语料库进行，具体的方法主要有：基于规则的方法，基于统计的方法和基于分类的方法。基于分类的方法目前研究的比较多，McCarthy 等人将判断先行语的问题转换成分类问题，通过分类器判断指代语与每个先行语候选之间是否存在指代关系；Soon 等人首次给出其详尽完整的实现步骤[2]，并开发出了实用的系统，在 MUC-6 上达到 *F*=62.6 的结果；Ng 等人对上述研究进行扩充[3]，抽取 53 个不同的词法、语法和语义特征，使消解的性能有所提高。之后的一些机器学习的方法多是对某种类别的名词短语进行消解，如文献[4]的研究主要是针对代名词。

本文实现一个基于支持向量机(Support Vector Machine, SVM)的英文名词短语指代消解的原型系统。SVM 高性能的分类效果已在许多领域被成功应用，在指代消解中也具有较好的性能。

2 语料资源

与自然语言处理的其他技术相同，基于机器学习的指代消解需要标注好的语料资源。目前，MUC 和 ACE 是标注了指代关系的常用语料资源，为更好地与其他指代消解系统进行比较，本文使用 MUC 语料库。

本文用标注好的 MUC-6 语料库进行训练和测试，取 30 篇标注过的 dryrun 文档作为训练文档，MUC-6 训练文档约有 12 400 个词，其中标有指代关系的有 2 141 个词，能构成指代的有 1 644 对。在测试时，用 MUC-6 的 30 篇 Formal 测试文档，其中约有 13 400 个词。

3 SVM 分类方法

指代消解问题可理解为二元分类问题，即判断 2 个名词短语是否存在指代关系，如有关系，则为正例；如没有关系，则为负例。分类方法对判断一个实例的正负非常重要，指代

基金项目：国家自然科学基金资助项目(60673041)；国家“863”计划基金资助项目(2006AA01Z147)
作者简介：李艳翠(1982－)，女，硕士，主研方向：中文信息处理；杨 勇，硕士研究生；周国栋，教授、博士生导师；朱巧明，教授
收稿日期：2008-05-15 **E-mail：**yancuili@gmail.com

消解常用的分类方法有 Decision Tree(C5.0, C4.5), RIPPER, MaxEn 和 SVM 等。

本文实验采用 SVM 进行分类，主要因为：(1)SVM 以统计学习理论为基础；(2)针对有限样本情况的，其目标是得到现有信息下的最优解；(3)算法将实际问题通过非线性变换转换到高维的特征空间(feature space)，可有效克服“维数灾难”；(4)人为设定的参数少，便于使用。

SVM 的主要思想是建立一个超平面作为决策曲面，使正例和负例之间的隔离边缘被最大化，SVM 基本原理如图 1 所示。

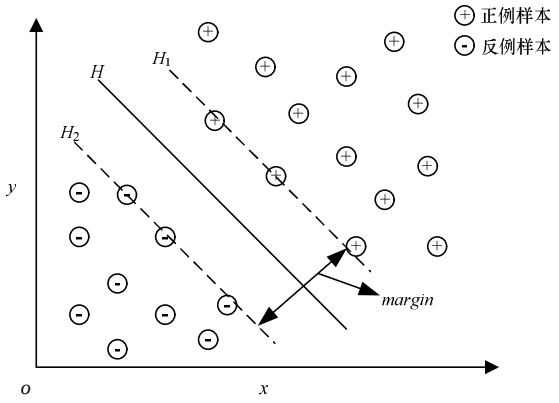

图 1　SVM 基本原理

其中，H为分类线，H_1, H_2分别代表各类中距离分类线最近且平行于分类线H的直线的样本，它们之间的距离称为分类间隔$margin=2/\|\boldsymbol{W}\|$，$\boldsymbol{W}$为最优超平面的法向量。

所谓最优分类线就是要求分类线能将 2 类样本正确分开且分类间隔最大。同理，如果在多维空间训练数据能被一个超平面没有错误地分开，并且离超平面最近的向量与超平面之间的距离最大，则称该平面为最优分类面[4]。

设有 n 个样本集 x_i 及其所属类别 y_i ，表示为

$x_i \in R^n$,　$y_i \in \{1,-1\}$,　$i=1,2,\cdots,n$

超平面 $w\cdot x+b=0$ 方程能将 2 类样本分开，即

$$y_i[(w\cdot x_i)+b]-1\geqslant 0 \qquad i=1,2,\cdots,n \tag{1}$$

当 $margin=2/\|w\|$, $\boldsymbol{H}_1$, $\boldsymbol{H}_2$ 上的训练样本点就称为支持向量。利用 Lagrange 优化方法得到最优分类函数为

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{n} a_i^* y_i (x_i\cdot x)+b^*\} \tag{2}$$

对于非线性问题，可通过非线性变换转化为某个高维空间中的线性问题，在变换空间求最优分类面，式(2)变为

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{n} a_i^* y_i K(x_i\cdot x)+b^*\}$$

其中，$K(x, x_i)$为核函数，利用此函数即可实现从低维空间向高维空间的映射，从而实现某一非线性分类变换为线性分类，a_i为样本对应的拉格朗日乘子，$y_i \in \{1,-1\}$。

除了线性核函数外，常用的核函数还有多项式核函数、径向基函数和 Sigmoid 函数。不同的分类核函数对分类结果也有所影响，本文用的 SVM 分类器是 Joachims 编写的二元分类器 SVMlight。

4　系统构架

本文采用较常用的基于语料库的机器学习方法进行名词短语的指代消解方法。训练语料经过预处理、特征向量的选择、特征值获得、基于规则过滤、训练实例生成、SVM 训练后生成分类模型文件，该模型是下文测试分类的依据。测试文本同样也要经过预处理、特征值抽取、基于规则过滤、配对名词短语形成测试实例，最后根据 SVM 给出的分类预测结果得出指代关系，系统框架如图 2 所示。

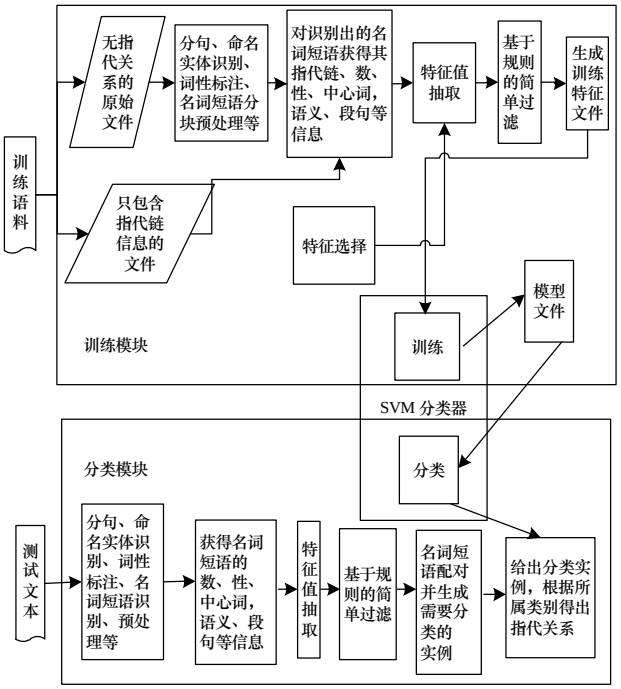

图 2　基于 SVM 的指代消解系统框架

4.1　预处理

自然语言文本进入系统之前，首先要进行预处理，为指代消解创造必要条件。预处理包括分句、命名实体的识别、词性标注、名词短语的识别、名词短语中心词的获得、语义信息的获得等过程。本文用到的预处理系统是文献[5-6]中的预处理系统，其中命名实体识别、词性标注和名词短语识别都是基于错误驱动的 HMM 模型；命名实体识别以及他们的语义类别是基于 HMM 的命名实体识别，在 MUC NE 任务中获得较高的 F 值，分别为 96.9%(MUC-6)和 94.3%(MUC-7)；名词短语的中心词获得用的是 Collins 的方法；语义信息是从 WordNet 中获得的。

在训练阶段，预处理前先将指代链抽取到一个文件中(如 MUC6 训练语料指代链上共有名词短语 2 141 个，其中能构成 1 644 个正例)，过滤掉训练文件标注的指代信息后，进行预处理，预处理完成后，再将处理生成的名词短语与读取的指代信息进行匹配(如 MUC6 指代链上名词短语正确匹配 1 978 个，匹配成功率为 91.4%)，然后再将根据名词短语指代信息生成正负例进行训练(训练时能产生 1 463 个正例，6 098 个负例)。测试阶段用的文本不包含指代信息，可直接进行预处理。预处理完成后须选择合适的特征向量交给 SVM 进行学习。

4.2　特征向量的选择

构建一个基于机器学习的指代消解系统时，选择合适的特征向量非常重要。特征向量的属性不能互相冲突，并且属性要有明确的表示。

属性的选择要具有一定的通用性，本文使用的特征是参照 Soon 等人于 2001 年提出的 12 个特征，SVM 只接受数值特征。实验的具体取值如表 1 所示。

表 1　指代消解系统所用到的特征及说明

特征	说明
ANPronoun	照应语是代词取 1，否则取 0 代词包括第一人称、第二人称、第三人称、中性代词、反身代词
ANDefiniteNP	照应语是有定名词短语取 1，否则取 0 有定是指以 The 开头的词
ANDemonstrativeNP	照应语是指示性名词短语取 1，否则取 0 指示是指以 That，This，These 或 Those 开头的词
CAPronoun	先行语是代词取 1，否则 0
ANCAGenderAgreement	照应语和先行语候选性别一致取 1，不一致取 0，其中 1 个不确定取 0.5。词的性有 Male，Female 或不确定
ANCANumberAgreement	照应语和先行语满足单复数一致取 1，不一致取 0，词的数有单数和复数，如代词 They 就是复数，Him 则是单数
ANCAAppositive	照应语和先行语是同位语取 1，否则取 0
ANCANameAlias	照应语和先行语是别名关系取 1，否则取 0 如 IBM 和 International Business Machines 是别名关系
ANCAFullStringMatch	如果照应语和先行语满足全匹配取 1，否则取 0 本文全匹配是指名词短语去除前面的 a, an, the 和 this, that 等词后全部转换为大写再进行匹配
ANCASentDistance	照应语和先行语在 1 句内取 1，2 句 0.9，……大于 10 句取 0
ANCAWordSense	从 Word Net 中获得的语义信息类有相同的取 1，无相同的取 0
ANCABothProperName	照应语和先行语候选全是专用名词取 1，否则取 0

4.3　训练样例和测试样例的产生

(1)构建训练实例:对于一个待消解的照应语(标注有指代关系的指向其他词的词)，要选择一定范围内的名词短语作为其候选项，本文采用 Soon 等人的方法，取照应语和先行语之间的词作为先行语候选集合 *CandidateSet* 包括{c_1, c_2,…, c_n}(先行语候选要经过一个简单的过滤，满足条件的才出现在在候选项集合中)，其中 $j>i$ 说明 c_j 比 c_i 更靠近照应语 *anaphor*。假设 *antecedent* 是 *anaphor* 指代的先行语，则 *NagativeSet = CandidateSet-antecedent*。将先行语和照应语组成的向量称为正例+*instance*(*antecedent, anaphor*)，其他名词短语与照应语构成负例-*instance*(c_i, *anaphor*)，其中，$c_i \in$ *NagetiveSet*。

(2)构建测试实例：测试文本预处理后得到名词短语，对每一个待消解的名词短语，与前面每一个名词短语依次构成测试实例，过滤后的实例交给分类器进行分类，每一个实例返回一个分类预测值，直到找到分类结果为正例。

5　实验结果

为方便与其他系统进行对比，本系统采用国际上通用的 MUC 评测程序进行评测。MUC 对指代消解结果的技术评估有 3 个重要标准：召回率 *R*, 准确率 *P* 和 *F* 值。其中，召回率 *R* 是指代消解结果中正确的对象数目占消解系统应消解的对象总数的百分比，它反映指代消解系统的完备性；准确率 *P* 是指代消解结果中正确的对象数目占实际消解的对象数目的百分比，它反映指代消解系统的准确程度。当比较 2 个不同指代系统的性能时，一般使用这 2 个指标的综合值 *F*:

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

其中，*P* 为准确率；*R* 为召回率；β 为召回率和准确率的相对权重，一般取 1。

实验表明，本系统的结果优于同类型的其他系统。具体实验结果见表 2。指代消解平台都选取了同样的特征，但选用的分类器是不一样的。其中，Soon 等人的指代消解平台采用 C5.0 作为分类器，Ng 等人采用 C4.5 和 Ripper 分类器，本文系统则采用 SVM。同时本文系统也选用其他的分类器进行了测试，实验结果见表 3。

表 2　本系统及其他同类系统结果比较

	MUC-6		
	R/(%)	*P*/(%)	*F*
SVM（径向基函数）	58.9	82.2	68.6
C5.0	58.6	67.3	62.6
C4.5/Ripper	62.4	70.7	66.3

表 3　不同分类器的分类结果

不同分类器	MUC-6		
	R/(%)	*P*/(%)	*F*
SVM(径向基函数)	58.9	82.2	68.6
C4.5	57.9	77.9	66.4
Ripper	46.6	82.7	59.7
MaxEnt	56.1	75.7	64.4

从表 3 中可以看出，本系统得到的基于 C4.5 的分类结果和 Ng 等人采用的方法中的 *F* 值基本相同，而 SVM 的结果明显优于 C4.5。说明在相同特征的条件下，SVM 确实能起到更好的分类效果。事实上，SVM 提供了多种核函数，而不同的核函数所得到的分类结果也不一样。表 4 给出了使用线性核函数、多项式核函数(*d*=2)、径向基函数(RBF)的对比结果。首先选取最重要的 3 个特征(同位语+别名+全匹配)进行系统训练和测试，单个特征的结果记录在表 5 中。实验表明，在 MUC 的指代消解任务中，使用径向基函数的效果最好。

表 4　SVM 分类器采用不同分类函数和特征的结果

SVM 函数类别	全部特征			不用距离特征			同位语+别名+全匹配		
	R/(%)	*P*/(%)	*F*	*R*/(%)	*P*/(%)	*F*	*R*/(%)	*P*/(%)	*F*
线性核函数	53.3	81.3	64.4	53.3	81.3	64.4	53.3	81.3	64.4
多项式核函数(*d*=2)	57.5	83.6	68.1	57.4	83.4	68.0	53.3	81.3	64.4
径向基函数(RBF)	58.9	82.2	68.6	59.0	82.2	68.7	53.3	81.3	64.4

表 5　只用一个特征的情况

	Soon 等人使用的系统			本文系统		
	R/(%)	*P*/(%)	*F*	*R*/(%)	*P*/(%)	*F*
ANPronoun	0.0	0.0	0.0	0.0	0.0	0.0
ANDefiniteNP	0.0	0.0	0.0	0.0	0.0	0.0
ANDemonstrativeNP	0.0	0.0	0.0	0.0	0.0	0.0
CAPronoun	0.0	0.0	0.0	0.0	0.0	0.0
ANCAGenderAgreement	0.0	0.0	0.0	0.0	0.0	0.0
ANCANumberAgreement	0.0	0.0	0.0	0.0	0.0	0.0
ANCAAppositive	3.9	57.7	7.3	3.7	67.9	7.0
ANCANameAlias	24.5	88.7	38.4	32.9	92.6	48.6
ANCAFullStringMatch	45.7	65.6	53.9	43.5	81.3	56.7
ANCASentDistance	0.0	0.0	0.0	0.0	0.0	0.0
ANCAWordSense	0.0	0.0	0.0	0.0	0.0	0.0
ANCABothProperName	0.0	0.0	0.0	0.0	0.0	0.0

由表 4 和表 5 可见，同位语、别名和全匹配 3 个特征对消解非常重要，尤其是全匹配特征，本系统达到召回率 43.5%，准确率 81.3%和 *F* 值 56.7，贡献度最为明显；其次为别名特征和同位语特征。这一结果与 Soon 等人的决策树方法相比，占有一定的优势。说明 SVM 在利用单个特征进行分类时仍能得到较好结果。

在实验中同时也发现，利用 SVM 进行分类时，并非所有的特征都能够起到较好的作用，如距离特征，若采用照应语和先行语候选的距离差的绝对值作为特征值(1 句内取 0，距离差为 1 时取 1，依此类推)，对 SVM 而言，只能起到干扰的作用，此时的 *F* 值只能达到 64.4。所以本系统的 F 值 68.6 是对距离特征进行加权处理的(同一句内取 1，距离为 1 取 0.9，大于 10 句取 0)，但是如果去掉距离特征，实验结果不会发生改变，如表 4 所示。

本系统虽然得到较好的准确率和 *F* 值，但与 Ng 等人提出的系统结果相比，召回率并没有优势，主要原因是本系统

（下转第 204 页）