基于测度优化 Laplacian SVM 的中文指代消解方法

周炫余[1,2],刘 娟[1,2],邵 鹏[1,2],卢 笑[3],罗 飞[1,2]
(1. 武汉大学软件国家重点实验室,湖北武汉 430072;2. 武汉大学计算机学院,湖北武汉 430072;
3. 湖南大学电气与信息工程学院,湖南长沙 410082)

摘 要: 相比于传统的基于半监督学习的指代消解方法,Laplacian SVM(Support Vector Machine)能有效的挖掘已标注样本和未标注样本的相似性和关联性,更好的推导模型的分类边界. 而传统 Laplacian SVM 采用欧式距离度量样本之间的距离,使得异类样本之间的相似性可能过大,不利于样本的准确分类. 对此,提出一种基于数据驱动学习最优测度 Laplacian SVM 算法以解决中文指代消解语料不足的问题. 该方法通过优化样本对之间的相似性约束条件和引入 Fisher 判别项,增大同类样本间的相似性,并突出强判别能力的特征. 此外,提出核嵌入的测度优化方法将以上线性测度优化推广到非线性空间,有利于 Laplacian SVM 利用核函数实现非线性分类. 在 ACE2005 中文语料库上的测评结果表明,所提出测度优化的 Laplacian SVM(包括线性和核嵌入两种形式)的方法只需少量标注样本就可以获得与经典的有监督学习模型相当甚至更好的消解性能,同时也优于其他传统的半监督学习方法.

关键词: 测度优化; Laplacian SVM; 中文指代消解; 半监督学习; 自然语言处理
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2016)12-3064-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.12.035

Chinese Anaphora Resolution Based on Metric-optimized Laplacian SVM

ZHOU Xuan-yu[1,2], LIU Juan[1,2], SHAO Peng[1,2], LU Xiao[3], LUO Fei[1,2]
(1. *State Key Laboratory of Software Engineering, Wuhan University, Wuhan, Hubei* 430072, *China*;
2. *Computer School, Wuhan University, Wuhan, Hubei* 430072, *China*;
3. *College of Electrical and Information Engineering, Hunan University, Changsha, Hunan* 410082, *China*)

Abstract: Compared to the traditional semi-supervised based anaphora resolution methods, Laplacian SVM (Support Vector Machine) can efficiently explore the similarity and correlations between labeled and unlabeled samples for deriving more accurate classification model. However, traditional Laplacian SVM simply uses Euclidean distance to calculate the distance between two samples, which may result that two samples from different classes may have false high similarity. To address the problem of insufficient Chinese annotated corpus, a data-driven based method is proposed to learn the optimal distance metric. The proposed method takes similarity constraints between sample-pairs into consideration and introduces the Fisher discrimination criterion, so that the similarities of in-class samples are higher than those of between-class samples, and the discriminant features are highlighted in the new metric space. Furthermore, the proposed metric-optimized method is generalized from linear to nonlinear space by the use of kernel, so that it can be used for non-linear classification. Compared with the classical supervised method and other four traditional semi-supervised methods on the ACE2005 Chinese corpus, the proposed method, both the linear form and kernel form, achieves the comparatively better or best performance, with fewer labeled samples.

Key words: metric-optimized; Laplacian SVM; Chinese anaphora resolution; semi-supervised learning; natural language processing

收稿日期:2015-11-05;修回日期:2016-06-30;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61272274);国家自然科学基金青年项目(No. 61402340);湖北省自然科学基金(No. 2014CFB194)

1　引言

指代消解是将篇章中指向同一实体(Entity)的实体表达(Mention)关联起来的过程,也是其余各类自然语言处理技术的关键子技术之一,例如:机器翻译[1]、自动问答[2]、自动文摘[3]、信息抽取[4]等. 基于机器学习的指代消解方法主要基于 Soon 等[5]提出的实体表达对模型(Mention-Pair Model),即先对实体表达对分类,后将指向同一实体的实体表达对进行聚类. 相比于英文,中文指代消解的标注语料较少且标注中文语料需要较高的专业文法知识. 如何利用有限的标注语料有效的处理中文指代消解问题显得尤为重要. 一些经典的半监督学习方法已经应用于指代消解领域[6-10],但这些方法都基于一个假设,即未标注样本的类标签由标注样本训练得来的分类模型所标定,这样的分类模型忽略了样本之间(包括已标注样本和未标注样本)的关联性和相似性. 当标注样本较少时,分类模型不能很好地挖掘未标注样本的相似性,用于推导模型的类别边界.

Laplacian SVM[11](以下简称 LapSVM)能充分挖掘样本之间的相似性,但它的性能很大程度上取决于样本间相似性的度量[12]. 在传统 LapSVM 中,连接样本间边的权重是基于欧式距离构造的,并不能凸显某些对分类具有较强判别能力的特征. 基于上述原因,本文提出一种测度优化的 LapSVM 用于中文指代消解. 利用现有的标注语料学习新的映射矩阵,将样本映射到其它便于分类的空间. 为了更好地挖掘未标注样本的信息,本文同时考虑样本对之间的相似性约束条件和样本的散列度,得到关于测度的目标函数,并且推导了相应的求解方法. 此外,提出该线性测度优化的核扩展方法,使其能方便地应用于核 LapSVM 中. 在 ACE2005 中文语料上采用五种方式对方法进行测评,结果表明本文所提出的基于测度优化的 LapSVM 算法只需少量标注样本就可以获取比有监督学习模型相当甚至更好的消解性能,同时也优于其他传统的半监督学习方法.

2　基于半监督学习的中文指代消解模型

基于半监督的中文指代消解模型由预处理、待消解项识别、指代消解、测评模块组成,具体如图 1 所示.

2.1　预处理模块

模型的预处理模块主要包括分词、词性标注、命名实体识别、句法分析、名词短语获取等. 在预处理过程中存在着大量的噪声,为了公平的探讨指代消解模型性能的好坏,文章将待消解项分为自动获取的实体表达(AutoMention)和标准实体表达(GoldMention).

2.2　待消解项识别模块

Stoyanov 等[13]在论文中指出待消解项识别是影响指代消解的关键因素之一,因此需要在指代消解模块前增加一个待消解项模块[14]. 该模块是基于层次过滤模型的思想,采用逐层过滤的方式获取最终的待消解项. 它主要分为扩充阶段和过滤阶段,扩充阶段为了尽可能提高待消解项识别的召回率,过滤阶段是在保证召回率不大幅下降的同时尽可能提高模块的准确率.

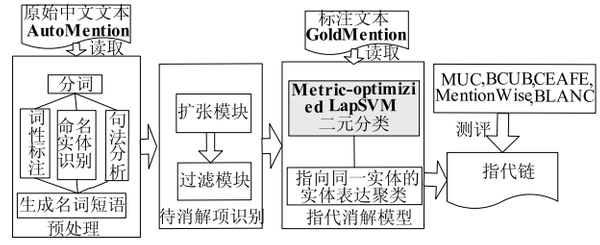

图1　基于半监督学习的中文指代消解框架

2.3　特征属性

本文将特征属性集合分为词法特征、语法特征、语义特征、位置特征四大类. 上述四类特征已经在前人的工作中证明了对指代消解是有效的,具体特征描述如表 1 所示.

表 1　特征属性集合

特征类别	特征	特征描述
词法特征	字符串匹配	如果照应语和先行语之间存在字符串匹配则为 1,否则为 0.
	宽松字符串匹配	如果照应语和先行语之间的字符串存在词包含关系时则为 1,否则为 0.
	中心词匹配	如果照应语和先行语之间存在中心词匹配则为 1,否则为 0.
	别名匹配	如果照应语和先行语之间存在别名匹配则为 1,否则为 0.
语法特征	同位语特征	如果照应语和先行语之间的是同位语为 1,否则为 0.
	生命特征	如果照应语和先行语之间的生物属性一致为 1,否则为 0.
	性别特征	如果照应语和先行语之间的性别一致为 1,否则为 0.
	单复数特征	如果照应语和先行语之间的单复数一致为 1,否则为 0.
语义特征	Web 语义知识	如果照应语和先行语之间的 Web 语义相似度小于 0.5 则二者为 1,否则为 0.
	本地语义知识 HowNet	如果照应语和先行语之间的 HowNet 语义相似度大于 0.5 则二者为 1,否则为 0.
位置特征	距离特征	如果照应语和先行语在一个句子中为 1,在第二个句子为 0.8,以此类推.

3　测度优化的 LapSVM 算法

3.1　传统 LapSVM 模型

给定 l 个已标注样本集合 $\{\boldsymbol{x}_i, y_i\}_{i=1}^{i=l}$ 和 u 个未标注的样本集合 $\{\boldsymbol{x}_j\}_{j=l+1}^{j=l+u}$,$y_i \in \{-1, +1\}$ 是标注类别信息,+1 样本存在指代关系,-1 则为样本不存在指代关系.

根据表 1 将样本抽象为 m 维向量,将决策函数表示为 f,传统的 LapSVM 求解模型如式(1)所示:

$$f^* = \arg\min \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_H \|f\|_H^2 + \frac{\gamma_M}{(l+u)^2} \boldsymbol{f}^{\mathrm{T}} \boldsymbol{L} \boldsymbol{f} \tag{1}$$

其中 V 是 hinge 损失函数 $\max[0, 1 - y_i f(x_i)]$;第二项是为了在重构希尔伯特空间(RKHS)找到最大分类间隔的光滑边界引入的正则项;第三项是为了利用流形的结构信息,而引入的流行正则项,以确保相似的样本有相同的类标. γ_H 和 γ_M 起到平衡后两项的作用. 此外,$\boldsymbol{L}$ 是图拉普拉斯矩阵,$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$,$\boldsymbol{W}$ 是所有数据邻接图的边的权重,传统 LapSVM 中,邻接图权重由欧式距离定义如式(2)所示:

$$w_{ij} = \exp\left(- \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\mathrm{F}}^2}{\sigma^2} \right) \tag{2}$$

$\|\cdot\|_{\mathrm{F}}$ 是 Frobenious 范数,对角矩阵 $\boldsymbol{D}$ 的各对角元素表示为 $\boldsymbol{D}_{ii} = \sum_{j=1}^{l+u} w_{ij}$,向量 $\boldsymbol{f}$ 是所有样本的决策函数值,$\boldsymbol{f} = [f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_{l+u})]^{\mathrm{T}}$. 根据图拉普拉斯矩阵的定义,式(1)中的第三项可以表示为

$$\begin{aligned} \boldsymbol{f}^{\mathrm{T}} \boldsymbol{L} \boldsymbol{f} &= \boldsymbol{f}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{f} - \boldsymbol{f}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{f} \\ &= \frac{1}{2} \sum_{i,j=1}^{l+u} w_{ij} (f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2 \end{aligned} \tag{3}$$

由式(3)及权重 w_{ij} 的定义可知,数据 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 在空间中的距离越小,则 w_{ij} 越大,必然使得两者的决策函数值越相近,或者说两者存在指代关系的概率越大;反之亦然. 因此,衡量数据的空间距离的测度对于模型起着至关重要的作用.

3.2 线性测度优化的 LapSVM 模型

考虑学习线性映射矩阵 $\boldsymbol{A}$,将样本映射到更具有判别能力的测度空间,此时样本的距离(马氏距离)如式(4)所示:

$$d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} (\boldsymbol{x}_i - \boldsymbol{x}_j) \tag{4}$$

由式(2)可知,w_{ij} 受半径参数 σ 影响较大,为避免 σ 的取值,重新定义样本间的权重如式(5)所示:

$$\begin{aligned} w_{ij}^* &= \exp(- \|\boldsymbol{A}(\boldsymbol{x}_i - \boldsymbol{x}_j)\|_{\mathrm{F}}^2) \\ &= \exp(- (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)) \end{aligned} \tag{5}$$

3.2.1 样本对相似性约束

对于已标注的样本,定义集合 S 和 F 分别表示属于同类的样本对和不属于同类的样本对,即:

$$S = \{(i,j) \mid \boldsymbol{x}_i \text{ 和 } \boldsymbol{x}_j \text{ 属于同一类}\},$$

$$F = \{(i,j) \mid \boldsymbol{x}_i \text{ 和 } \boldsymbol{x}_j \text{ 属于不同类}\}.$$

测度优化的目的是增大同类样本对之间的相似性,减小异类样本对之间的相似性,因此,考虑如式(6)所示的相似性约束优化线性矩阵 $\boldsymbol{A}$:

$$S(\boldsymbol{A}) = \sum_{(i,j)\in F} w_{ij}^* - \sum_{(i,j)\in S} w_{ij}^* \tag{6}$$

为了防止过拟合,在目标函数中引入所有样本权重之和以保证权重变化在一定范围内,得到目标函数如式(7)所示:

$$S(\boldsymbol{A}) = \sum_{(i,j)\in F} w_{ij}^* - \sum_{(i,j)\in S} w_{ij}^* - \mu \sum_{i=1}^{l+u} \sum_{j\in N_j} w_{ij}^* \tag{7}$$

其中 μ 是平衡因子,N_j 是样本 $\boldsymbol{x}_j$ 的 k 最近邻.

3.2.2 样本 Fisher 判别项

为了突出判别能力强的特征更好地应用于分类,本文引入 Fisher 判别项[15],如式(8)所示:

$$F(\boldsymbol{A}) = \mathrm{tr}\,(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{A} - \boldsymbol{A}^{\mathrm{T}} \boldsymbol{S}_B \boldsymbol{A}) \tag{8}$$

其中 $\mathrm{tr}(\boldsymbol{A})$ 表示矩阵 $\boldsymbol{A}$ 的迹,$\boldsymbol{S}_w$ 是类内散列矩阵,$\boldsymbol{S}_B$ 是类间散列矩阵,分别如式(9)(10)所示:

$$\boldsymbol{S}_w = \sum_{i\in(-1,1)} \sum_{y_k = i} (\boldsymbol{x}_k - \boldsymbol{m}_i)(\boldsymbol{x}_k - \boldsymbol{m}_i)^{\mathrm{T}} \tag{9}$$

$$\boldsymbol{S}_B = \sum_{i\in(-1,1)} n_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^{\mathrm{T}} \tag{10}$$

其中 n_i 是正样本或者负样本的个数,$\boldsymbol{m}_i$ 是正样本或者负样本的均值,$\boldsymbol{m}$ 是所有标注样本的均值. 另外,引入关于 $\boldsymbol{A}$ 的二次项,$\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} - \boldsymbol{I}\|_{\mathrm{F}}^2$ 控制其复杂度.

3.2.3 线性测度优化目标函数及求解方法

综上所述,线性测度优化的目标函数可用式(11)表示,其中 λ、β 分别为平衡因子.

$$\begin{aligned} Q(\boldsymbol{A}) &= S(\boldsymbol{A}) + \lambda F(\boldsymbol{A}) + \beta \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} - \boldsymbol{I}\|_{\mathrm{F}}^2 \\ &= \sum_{(i,j)\in F} w_{ij}^* - \sum_{(i,j)\in S} w_{ij}^* - \mu \sum_{i=1}^{l+u} \sum_{j\in N_j} w_{ij}^* \\ &\quad + \lambda \mathrm{tr}\,(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{A} - \boldsymbol{A}^{\mathrm{T}} \boldsymbol{S}_B \boldsymbol{A}) + \beta \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} - \boldsymbol{I}\|_{\mathrm{F}}^2 \end{aligned} \tag{11}$$

为避免问题(11)中对 $\boldsymbol{A}$ 求导的繁琐,令 $\boldsymbol{M} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$,可直接求解 $\boldsymbol{M}$,$\boldsymbol{M}$ 是半正定矩阵,根据式(5),对 $\boldsymbol{M}$ 求导的结果如式(12)所示:

$$\frac{\partial w_{ij}^*}{\partial \boldsymbol{M}} = \sum_{(i,j)\in F} - w_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \tag{12}$$

则:

$$\begin{aligned} \frac{\partial S(\boldsymbol{M})}{\partial \boldsymbol{M}} &= \sum_{(i,j)\in F} - w_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \\ &\quad - \sum_{(i,j)\in S} - w_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \\ &\quad - \mu \sum_{i=1}^{l+\mu} \sum_{j\in N_j} - w_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \end{aligned} \tag{13}$$

且 $\frac{\partial F(\boldsymbol{M})}{\partial \boldsymbol{M}} = \boldsymbol{S}_W - \boldsymbol{S}_B$,则目标函数 $Q(\boldsymbol{M})$ 对 $\boldsymbol{M}$ 求导如式(14)所示:

$$\frac{\partial Q(\boldsymbol{M})}{\partial \boldsymbol{M}} = \frac{\partial S(\boldsymbol{M})}{\partial \boldsymbol{M}} + \lambda \frac{\partial F(\boldsymbol{M})}{\partial \boldsymbol{M}} + 2\beta(\boldsymbol{M} - \boldsymbol{I}) \tag{14}$$

将 $\boldsymbol{A}$ 的初始值设置为 n 行 m 列的随机矩阵,则 $\boldsymbol{M}$ 的初值为 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$,采用梯度下降法求解式(14)的最优值,迭代更新原则如式(15)所示:

$$\boldsymbol{M}_{t+1}=\boldsymbol{M}_t-\eta_t\frac{\partial Q}{\partial \boldsymbol{M}_t} \tag{15}$$

其中步长 η_t 设置采用 backtracking line search 的方法.

所提出的线性测度优化 LapSVM 方法的算法流程如算法 1 所示.

算法 1　基于线性测度优化的 LapSVM 算法

输入:l 个已标注样本集合 $\{\boldsymbol{x}_i,y_i\}_{i=1}^{i=l}$ 和 u 个未标注的样本集合 $\{x_j\}_{j=l+1}^{j=l+u}$;

输出:决策函数 f^*;

1. 初始化 $\boldsymbol{A}$,将 $\boldsymbol{A}$ 设置为 n 行 m 列的随机矩阵,$\boldsymbol{M}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$,设置 $t=0$;
2. 使用式(5)构造一个图拉普拉斯矩阵边的权重 w_{ij}^*;
3. 利用式(12)、(13)和(14)计算 $Q(\boldsymbol{M})$和$\partial Q(\boldsymbol{M})/\partial \boldsymbol{M}$;
4. 通过式(15)计算 $\boldsymbol{M}_{t+1}=\boldsymbol{M}_t-\eta_t(\partial Q(\boldsymbol{M})/\partial \boldsymbol{M})$,如果 $Q(\boldsymbol{M}_{t+1})>Q(\boldsymbol{M}_t)$,$\eta_{t+1}=0.5\eta_t$ 且 $\boldsymbol{M}_{t+1}=\boldsymbol{M}_t$,否则 $\eta_{t+1}=2\eta_t$;
5. 若 $t>T$,则设置 $t=t+1$(T 是算法的迭代次数),退出迭代,输出矩阵 $\boldsymbol{M}$,否则转至步骤 2;
6. 对 $\boldsymbol{M}$ 做 SVD 分解,$\boldsymbol{M}^*=\boldsymbol{U}\boldsymbol{S}_+\boldsymbol{V}^{\mathrm{T}}$,其中 $\boldsymbol{S}_+$ 是将 $\boldsymbol{M}$ 的特征值矩阵中的小于零的值置零的结果;
7. 利用更新的 $\boldsymbol{M}^*$ 矩阵和式(5)得到新的图中边的权重 w_{ij}^*,并得到新的图拉普拉斯矩阵 $\boldsymbol{L}^*$;
8. 利用更新的 $\boldsymbol{L}^*$ 代替 $\boldsymbol{L}$ 求解新的 LapSVM 问题(1)得到决策函数 f^*.

3.3　核嵌入的测度优化的 LapSVM 模型

由于中文语料特征维数较少,将其映射到高维空间,或通过提取语料的结构化特征[16],进而利用相应的核函数将更有利于非线性数据的分类. 为了能够有效利用核函数,本节将上节提出的线性测度优化 LapSVM 扩展至非线性空间,即,核嵌入的测度优化的 LapSVM.

3.3.1　样本对相似性约束的核扩展

若非线性映射 $\varphi(\boldsymbol{x})$将样本 $\boldsymbol{x}$ 映射到特征空间 $\mathbf{H}$:$\boldsymbol{x}\rightarrow\varphi(\boldsymbol{x})\in\mathbf{H}$,其相应的核函数为 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j)\rangle$,$\mathbf{H}$ 的维度表示为 d(可能为无限维),则 $\mathbf{H}$ 中两数据之间的马氏距离,如式(16)所示:

$$\begin{aligned}&d_{\mathbf{H}}^2(\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j))\\&=(\varphi(\boldsymbol{x}_i)-\varphi(\boldsymbol{x}_j))^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}(\varphi(\boldsymbol{x}_i)-\varphi(\boldsymbol{x}_j))\\&=(\varphi(\boldsymbol{x}_i)-\varphi(\boldsymbol{x}_j))^{\mathrm{T}}\boldsymbol{M}(\varphi(\boldsymbol{x}_i)-\varphi(\boldsymbol{x}_j))\end{aligned} \tag{16}$$

根据相关的测度学习理论[17],核测度学习无需在 $\mathbf{H}$ 空间中显示地学习测度矩阵 $\boldsymbol{M}$,其最优参数化模型的形式为 $\boldsymbol{M}=\varphi(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\varphi(\boldsymbol{X})$,其中 $\boldsymbol{X}$ 表示 m 行 n 列的样本特征矩阵,因此根据式(16),$\mathbf{H}$ 中马氏距离定义为式(17)所示:

$$d_{\mathbf{H}}^2(\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{x}_j))=(\boldsymbol{K}^i-\boldsymbol{K}^j)^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}(\boldsymbol{K}^i-\boldsymbol{K}^j) \tag{17}$$

其中 $\boldsymbol{K}^i$ 是核矩阵 $\boldsymbol{K}$ 的第 i 列,其第 i 行 j 列元素为 $\boldsymbol{K}_{ij}=k(\boldsymbol{x}_i,\boldsymbol{x}_j)$. 因此,核嵌入下,权重 w_{ij}^* 可表示为式(18):

$$\begin{aligned}w_{ij}^{k*}&=\exp(-(\boldsymbol{K}^i-\boldsymbol{K}^j)^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}(\boldsymbol{K}^i-\boldsymbol{K}^j))\\&=\exp(-(\boldsymbol{K}^i-\boldsymbol{K}^j)^{\mathrm{T}}\boldsymbol{N}(\boldsymbol{K}^i-\boldsymbol{K}^j))\end{aligned} \tag{18}$$

利用式(18)可避免显示地使用非线性映射 $\varphi(\boldsymbol{x})$而直接利用核函数 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$求解映射矩阵 $\boldsymbol{P}$ 或直接求解 $\boldsymbol{N}$.

3.3.2　样本 Fisher 判别项的核扩展

根据文献[18]等提出的理论,Fisher 判别项的核扩展(Kernel Fisher Discriminant)如式(19)所示:

$$F_k(\boldsymbol{P})=\mathrm{tr}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{S}_w^k\boldsymbol{P}-\boldsymbol{P}^{\mathrm{T}}\boldsymbol{S}_B^k\boldsymbol{P}) \tag{19}$$

其中类内散列矩阵如式(20)所示:

$$\boldsymbol{S}_W^k=\sum_{j=1,-1}\boldsymbol{K}_j(\boldsymbol{I}-\mathbf{1}_{n_j})\boldsymbol{K}_j^{\mathrm{T}} \tag{20}$$

其中 $\boldsymbol{K}_j$ 的第 n 行 m 列元素定义为 $k(\boldsymbol{x}_n,\boldsymbol{x}_m^j)$,$I$ 是单位矩阵,$\mathbf{1}_{n_j}$是所有元素都为 $1/n_j$ 的向量. 类内散列矩阵如式(21)所示:

$$\boldsymbol{S}_W^k=(\boldsymbol{S}_1-\boldsymbol{S}_{-1})(\boldsymbol{S}_1-\boldsymbol{S}_{-1})^{\mathrm{T}} \tag{21}$$

其中$(\boldsymbol{S}_i)_{j(i=1,-1)}=1/n_i\sum_{k=1}^{n_i}k(\boldsymbol{x}_j,\boldsymbol{x}_k^i)$.

基于以上分析,核嵌入的测度优化与线性测度优化目标函数有相同的形式,为求解方便,可直接求解 $\boldsymbol{N}$,将 $\boldsymbol{N}$ 替换式(18)中的 $\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}$,得到 $Q^k(\boldsymbol{N})$,具体如式(22)所示:

$$\begin{aligned}Q^k(\boldsymbol{N})&=S^k(\boldsymbol{N})+\lambda F^k(\boldsymbol{N})+\beta\|\boldsymbol{N}-\boldsymbol{I}\|_{\mathrm{F}}^2\\&=\sum_{(i,j)\in F}w_{ij}^{k*}-\sum_{(i,j)\in S}w_{ij}^{k*}-\mu\sum_{i=1}^{l+u}\sum_{j\in N_i}w_{ij}^{k*}\\&+\lambda\mathrm{tr}(\boldsymbol{S}_W^k\boldsymbol{N}^{\mathrm{T}}-\boldsymbol{S}_B^k\boldsymbol{N}^{\mathrm{T}})+\beta\|\boldsymbol{N}-\boldsymbol{I}\|_{\mathrm{F}}^2\end{aligned} \tag{22}$$

其中各项对 $\boldsymbol{N}$ 求导如式(23)、(24)所示:

$$\begin{aligned}\frac{\partial S^k(\boldsymbol{N})}{\partial \boldsymbol{N}}=&\sum_{(i,j)\in F}-w_{ij}(\boldsymbol{K}^i-\boldsymbol{K}^j)(\boldsymbol{K}^i-\boldsymbol{K}^j)^{\mathrm{T}}\\&-\sum_{(i,j)\in S}-w_{ij}(\boldsymbol{K}^i-\boldsymbol{K}^j)(\boldsymbol{K}^i-\boldsymbol{K}^j)^{\mathrm{T}}\\&-\mu\sum_{i=1}^{l+\mu}\sum_{j\in N_i}-w_{ij}(\boldsymbol{K}^i-\boldsymbol{K}^j)(\boldsymbol{K}^i-\boldsymbol{K}^j)^{\mathrm{T}}\end{aligned} \tag{23}$$

$$\frac{\partial F^k(\boldsymbol{M})}{\partial \boldsymbol{M}}=\boldsymbol{S}_W^k-\boldsymbol{S}_B^k \tag{24}$$

则目标函数 $Q^k(\boldsymbol{N})$对 $\boldsymbol{N}$ 求导如式(25)所示:

$$\frac{\partial Q^k(\boldsymbol{N})}{\partial \boldsymbol{N}}=\frac{\partial S^k(\boldsymbol{N})}{\partial \boldsymbol{N}}+\lambda\frac{\partial F^k(\boldsymbol{N})}{\partial \boldsymbol{N}}+2\beta(\boldsymbol{N}-\boldsymbol{I}) \tag{25}$$

采用梯度下降法对 $\boldsymbol{N}$ 进行更新设置如式(26)所示:

$$\boldsymbol{N}_{t+1}=\boldsymbol{N}_t-\eta_t\partial Q^k/\partial \boldsymbol{N}_t \tag{26}$$

所提出的核嵌入测度优化 LapSVM 方法的算法流程如算法 2 所示.

算法 2　基于核嵌入的测度优化的 LapSVM 算法

输入:l 个已标注样本$\{\boldsymbol{x}_i,y_i\}_{i=1}^{i=l}$和 u 个未标注的样本$\{x_j\}_{j=l+1}^{j=l+u}$,核矩阵 $\boldsymbol{K}$,$\boldsymbol{K}_{ij}=k(\boldsymbol{x}_i,\boldsymbol{x}_j)$;

输出:决策函数 f^*;

1. 初始化 $\boldsymbol{P}$,将 $\boldsymbol{P}$ 设置为 n 行 $l+u$ 列的随机矩阵,$\boldsymbol{N}=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}$,设 $t=0$;
2. 使用式(18)构造图拉普拉斯矩阵边的权重 w_{ij}^{k*};
3. 利用式(22)~(25)计算 $Q^k(\boldsymbol{N})$ 和$\partial Q^k(\boldsymbol{N})/\partial \boldsymbol{N}$;
4. 通过式(26)计算 $\boldsymbol{N}_{t+1}=\boldsymbol{N}_t-\eta_t(\partial Q^k/\partial \boldsymbol{N})$,如果 $Q^k(\boldsymbol{N}_{t+1})>Q^k(\boldsymbol{N}_t)$,$\eta_{t+1}=0.5\eta_t$ 且 $\boldsymbol{N}_{t+1}=\boldsymbol{N}_t$,否则 $\eta_{t+1}=2\eta_t$;
5. 若 $t>T$,则设置 $t=t+1$(T 是算法的迭代次数),退出迭代,输出矩阵 $\boldsymbol{N}$,否则转至步骤2;
6. 对 $\boldsymbol{N}$ 做 SVD 分解,$\boldsymbol{N}^*=\boldsymbol{U}\boldsymbol{S}_+\boldsymbol{V}^{\mathrm{T}}$,其中 $\boldsymbol{S}_+$ 是将 $\boldsymbol{N}$ 的特征值矩阵中的小于零的值置零的结果;
7. 利用更新的 $\boldsymbol{N}^*$ 矩阵和式(4)得到新的图中边的权重 w_{ij}^{k*},并得到新的图拉普拉斯矩阵 $\boldsymbol{L}^{k*}$;
8. 利用更新的 $\boldsymbol{L}^{k*}$ 代替原有的 $\boldsymbol{L}$ 求解新的 LapSVM 问题(1)得到决策函数 f^*.

3.4 测度优化对分类器性能影响分析

从 ACE2005 中文语料中随机抽取 70 篇作为已标注训练语料,100 篇作为未标注训练语料,100 篇作为测试语料. 根据表 1 提取各实体表达对的 11 维特征. 对于核嵌入的测度优化算法,采用如式(27)所示的核映射:

$$k(\boldsymbol{x},\boldsymbol{y})=(\langle \boldsymbol{x},\boldsymbol{y}\rangle+1)^d(d=5) \tag{27}$$

通过各分类器的 ROC 曲线比较测度优化前、线性测度和核测度优化后分类器的性能,具体结果如图 2 所示. 其中横纵坐标分别表示分类结果的假阳率(FPR)和真阳率(TPR). 结果表明,线性测度优化和核测度优化后的分类器性能均明显优于测度优化前的性能. 这充分说明了通过映射矩阵 $\boldsymbol{A}$ 的空间变换,能增强特征的判别能力,从而改善分类器的性能,并且线性测度优化的核扩展进一步增强了分类器对于非线性数据的分类能力.

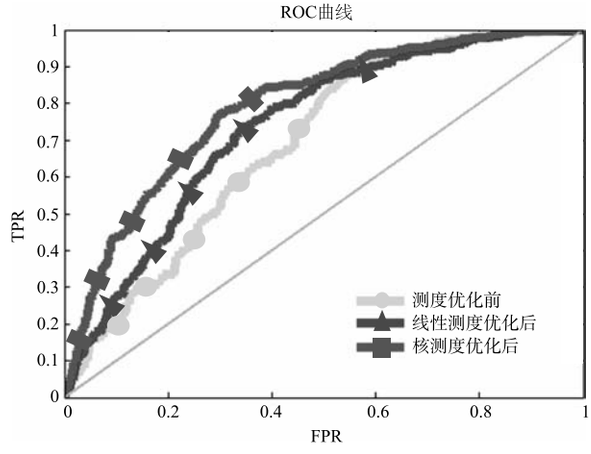

图2 测度优化前、线性测度优化后及核测度优化后分类器ROC曲线

4 实验与结果分析

4.1 实验设置

本文采用 ACE2005 中文语料为实验数据,并采用国际上通用的五种测评方式对指代消解模型进行测评. 为了更加公正的比较各类模型的指代消解精度,从 ACE2005 中文语料中的三个子语料中分别抽取出 45、43、12 篇组成开发集用于模型的参数调优,剩余的 533 篇作为训练集和测试集,并从训练集中随机抽取 200 篇作为半监督学习模型的标注语料,剩余的为未标注语料. 最后采用十倍交叉验证的方式对所有模型进行测评和比较,实验数据在三个子语料上的具体分布如表 2 所示.

表 2 实验语料统计表

语料库	总篇章数	BNEWS 篇数	Newswire 篇数	Weblog 篇数	实体总数	实体表达总数	实体表达/实体
训练集+测试集	533	253	195	85	12401	27745	2.24
开发集	100	45	43	12	3013	6576	2.18
总共	633	298	238	97	15414	34321	2.22

五类测评方法分别是 MUC[19]、BCUBED[20]、CEAFE[21]、MentionPairs[22]、BLANC[23]. 上述五类测评主要通过正确率 P(Precision)、召回率 R(Recall)、综合评定值 F 来衡量系统的消解性能.

4.2 对比模型

本节将取四类传统的半监督的学习方法和一类经典的有监督的学习方法作为对比模型,它们分别是 Kehler 等[6]提出 Self-Training 模型、Müller 等[7]提出 Co-Training 模型、Ng 等[9]提出的 EM 模型、Belkin 等[11]提出的传统的 LapSVM 算法以及 Soon 等[5]提出的实体表达对模型.

4.3 结论与错误分析

为了公平的比较各个模型的消解精度,在开发集上对所有的模型进行参数调优.

4.3.1 模型的参数设置

Self-Training 模型参数: Self-Training 模型参数只需要设置 bag 的个数. 为了获取最优的 bag 个数,分别测试了 bag 数为 1 到 20,结果表明当 bag 个数为 9 时可以取得最好的 F 值,具体实验结果如图 3 所示.

Co-Training 模型参数: 由于 Co-Training 模型参数众多,且该模型对参数比较敏感,因此为了降低实验的复杂度,根据前人的研究基础[7]设定一系列基础的参数,只对模型的迭代次数进行优化,具体优化如图 4 所示. 模型中正负样本比例设置为 2∶6;增长规模设置为 50. 在开发集上当迭代次数为 300 的时候取得最好消解能力.

EM 模型参数: EM 模型需要对迭代次数进行优

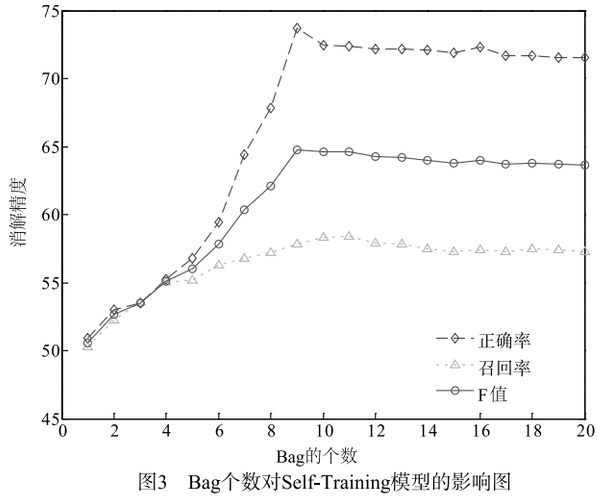

图3　Bag个数对Self-Training模型的影响图

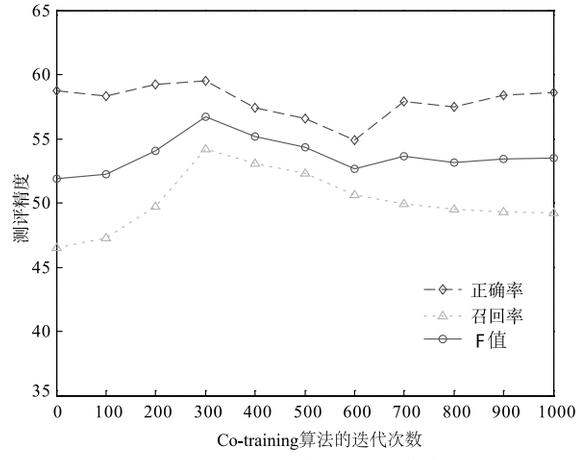

图4 Co-Training模型的学习曲线图

化. 模型分别测试的迭代次数从 5 至 30,结果表明当算法迭代至 20 次时算法可以取得最好的 F 值,具体结果如图 5 所示.

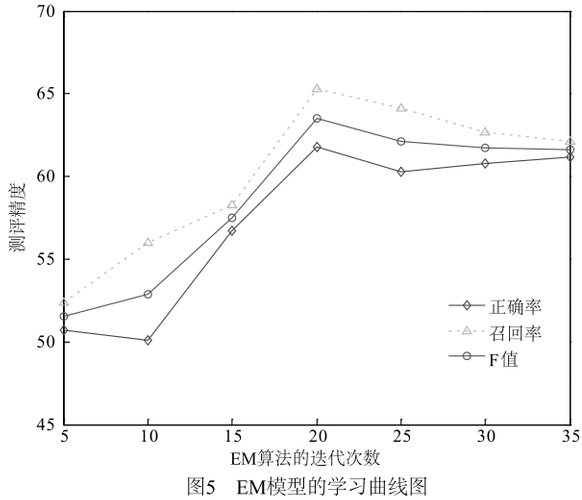

图5　EM模型的学习曲线图

传统的 LapSVM 模型和测度优化的 LapSVM 模型:对式(1)中的两个参数 γ_H 和 γ_M 在候选集中 $\gamma_H, \gamma_M \in \{10^{-3}, 10^{-2}, \cdots, 10^2, 10^3\}$ 进行网格搜索交叉验证,得到 $\gamma_H = 0.1, \gamma_M = 1e-4$. 其中基于核嵌入 LapSVM 采用如式(27)所示的五次多项式核函数.

4.3.2　AutoMention 与 GoldMention 的消解结果比较

从表 3 可以看出六种模型的 AutoMention 测评和 GoldMention 测评的 Avg 值存在着约 7% ~9% 的差距,主要有以下两个原因:

(1)中文预处理存在着一定程度上的错误. 例如:由于分词工具所采用的分词标准和标注语料不同,以及分词模型的自身不足导致中文分词一直是一个中文信息处理中悬而未决的问题. 指代消解模型的预处理模块是一个经典的管道模型,分词处理后续的词性标注、命名实体识别、句法分析等中文信息处理都是基于分词的结果进行处理的. 分词结果的好坏在一定程度左右着后续预处理的结果,例如:在具体试验中,分词工具将待消解项“文化局长”,“龙应台”分成了“文化局”、“长龙”、“应台”. 预处理不准确将极大的影响待消解项识别和指代消解模型的准确率和召回率.

(2)待消解项的识别模块仍旧不够准确. 表 4 显示了待消解项模块在 BNEWS、NWIRE、Weblogs 子语料上的测评结果,可见整体测评精度(ALL_AVG)偏低. 其主要原因是,待消解项模块为了更好的识别语料中的嵌套名词短语,从句法分析树上抽取包含 NN、NP、NR、EN 的节点作为候选待消解项,这样引入了许多中心词重复的待消解项.

4.3.3　测度优化的 LapSVM 算法与基于 SVM 的实体表达对模型的比较

实体表达对模型是目前使用最为广泛的指代消解模型,因此和该类模型进行比较具有一定的意义. 模型性能的高低通过五种测评方式的 F 值的平均值 Avg_F 来刻画. 表 3 中的 L_LapSVM 表示基于线性测度优化的 LapSVM 模型,K_LapSVM 表示基于核嵌入的测度优化的 LapSVM 模型. 由表 3 可知,实体表达对模型的 Avg 值为 60.2% 和 69.1%,基于线性测度优化的 LapSVM 模型 Avg 值为 59.6% 和 68.7%;基于核嵌入的测度优化的 LapSVM 模型 Avg 值为 62.8% 和 71.8%. 基于线性测度优化的 LapSVM 模型在综合性能上和实体表达对模型基本相当,但实体表达对模型所需的训练语料远多于线性测度优化的 LapSVM 模型,且基于核嵌入测度优化的 LapSVM 在训练语料较少的条件下,能获得比实体表达对模型更好的消解效果,这在标注语料不足的语种中具有显著的意义.

表 3 模型与 Baseline 模型的对比实验结果

待消解项类别	模型	ACE2005 中文语料															
		MUC			BCUB			CEAFE			MentionWise			BLANG			Avg_F
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	F
AutoMention	Self-Training	61.1	50.3	55.1	68.8	56.4	61.9	56.1	56.9	56.4	44.2	42.0	43.1	52.4	49.1	50.6	53.4
	Co-Training	51.2	42.7	46.5	60.2	54.6	57.2	47.9	53.0	50.3	42.4	40.5	41.4	46.0	44.2	45.1	48.2
	EM	50.4	56.8	53.4	62.1	60.8	61.4	52.4	55.9	54.0	45.3	45.8	45.5	45.7	46.7	46.1	52.1
	LapSVM	55.4	52.3	53.8	62.7	55.1	58.7	54.8	56.3	55.5	48	46.9	47.4	55.6	56.2	55.9	54.2
	Mention-Pair	60.1	59.8	59.9	72.1	62.8	67.1	58.0	60.2	59.0	55.3	51.6	53.3	65.7	58.3	61.7	60.2
	L_LapSVM	**61.4**	**58.6**	**60.0**	**72.8**	**62.1**	**67.0**	**57.3**	**63.4**	**60.2**	**52.4**	**46.2**	**49.1**	**64.1**	**58.8**	**61.3**	**59.6**
	K_LapSVM	**66.3**	**60.7**	**63.3**	**75.2**	**65.5**	**70.0**	**61.7**	**68.6**	**64.9**	**56.4**	**49.3**	**52.6**	**66.7**	**61.3**	**63.9**	**62.8**
GoldMention	Self-Training	71.8	55.2	62.4	74.3	60.4	66.6	62.7	63.8	63.2	59.5	43.1	50.0	66.4	58.9	62.4	60.8
	Co-Training EM	60.3	52.3	56.0	67.3	61.2	64.1	53.1	58.7	55.7	52.6	46.5	49.4	57.4	53.1	55.2	55.9
	EM	59.8	64.7	62.1	66.5	75.4	70.7	56.8	61.0	58.8	50.1	51.3	50.6	56.5	62.5	59.4	60.3
	LapSVM	60.1	62.0	61.0	69.3	69.7	69.4	57.2	66.9	61.6	55.6	50.7	53.0	67.8	67.4	68.4	62.4
	Mention-Pair	72.3	64.8	68.3	80.5	71.6	75.8	67.7	72.5	70.0	62.1	57.3	59.6	76.2	67.4	71.5	69.1
	L_LapSVM	**72.0**	**64.1**	**67.8**	**80.6**	**71.8**	**75.9**	**66.9**	**78.0**	**72.1**	**61.7**	**54.3**	**57.8**	**73.6**	**66.9**	**70.1**	**68.7**
	K_LapSVM	**74.6**	**67.3**	**70.8**	**83.2**	**76.4**	**79.7**	**70.6**	**80.1**	**75.1**	**64.3**	**55.4**	**59.5**	**76.5**	**71.8**	**74.1**	**71.8**

表 4 待消解项识别模块的性能表

待消解项识别模块	语料库	P	R	F
Our model	BNEWS(253 篇)	79.3	81.5	80.3
Our model	NWIRE(195 篇)	80.2	81.9	81.1
Our model	Weblogs(85 篇)	70.9	76.8	73.8
Our model	ALL_AVG(533 篇)	**78.3**	**80.9**	**79.6**

4.3.4 测度优化的 LapSVM 算法与其余半监督学习方法的比较

从表 3 可知,基于单视角的几类模型要好于基于多视角的 Co-Training 模型,主要原因有以下三点:(1)是 Co-Training 模型需要设置的参数较多,模型难以优化出合适的参数集合;(2)该模型的建立是基于特征选取角度相互独立的假设,然而这一假设在中文指代消解中是不成立的;(3)模型在参数发生变化时消解结果变化比较剧烈,模型对参数比较敏感. EM 模型的运行结果依赖于生成模型的正确性,然而文章中的生成模型是基于朴素贝叶斯的思想,朴素贝叶斯的前提条件是各特征之间是相互独立的. 指代消解的特征属性存在一定的冗余性,这显然与该前提条件存在一定的矛盾,因此 EM 模型的测评中结果仅仅只是好于 Co-Training 模型. Self-Training 模型虽然思想比较简单,但是该模型的正确率在所有的对比模型中比较高,主要是因为以下两点:(1)由图 5 可知,Self-Training 模型采用 bag 数为 9 好于 bag 数为 0,且随着 bag 数在 0 至 9 的区间逐渐增高,因此基于投票机制的 bagging 算法能提高 Self-Training 模型的消解精度;(2)Self-Training 模型在每一次迭代的过程中只将所有子分类器标记一致的样本加入到原始标注语料中,因此很大程度上保证了样本的可信度,尽量减少错误样本的加入. 然而,Self-Training 模型并未考虑整个数据空间的一致性,没有利用未标注样本的类结构信息,基于测度学习的 LapSVM 模型在考虑标注样本的同时,也充分的考虑了未标注样本之间的关联性和相似性. 该模型将标注样本和未标注样本融合进模型,从整个数据空间一致的角度来执行分类模型,因此该分类模型具有更好的消解效果. 从表 3 可知,在 GoldMention 中基于测度优化的 LapSVM 模型在综合 F 值分别好于 Self-Training 模型约 7.9% 和 11%. 基于线性测度优化学习的 LapSVM 模型优于传统的 LapSVM 模型,可见数据驱动学习得到的测度优于人工定义的测度,有利于 LapSVM 更好的挖掘样本的相似性推导最优分类函数. 此外,基于核嵌入测度优化的 LapSVM 模型使用核函数,实现了非线性空间的测度学习和分类函数优化,进一步增强了模型的分类能力,因此该模型的消解性能好于其余几类对比模型.

5 结论

提出一种基于测度优化 Laplacian SVM 算法的半监督学习方法,用于解决中文指代消解语料不足的问题. 建立了用于优化样本对之间相似性和样本间散列度的目标函数,并推导了最优测度的求解方法. 此外,核嵌入测度优化 LapSVM 能将分类函数优化与测度

优化相结合,并有效利用核函数实现非线性分类. 在 ACE2005 中文语料上进行测评,结果表明,所提出的方法能取得比基于 SVM 的实体表达对模型相当甚至更好的测评效果,且所需的训练样本更少;同时,指代消解效果也明显的优于其它四类传统的半监督学习方法.

参考文献

[1] HARDMEIER C, FEDERICO M. Modelling pronominal anaphora in statistical machine translation[A]. Proceedings of the International Workshop on Spoken Language Translation[C]. Paris: Konferensbidrag, 2010. 283 – 289.

[2] 张志昌,张宇,刘挺,等. 开放域问答技术研究进展[J]. 电子学报,2009,37(5):1058 – 1069.
ZHANG Zhi-chang, ZHANG yu, LIU Ting, et al. Advances in open-domain question answering[J]. Acta Electronica Sinica, 2009, 37 (5): 1058 – 1069. (in Chinese)

[3] 杨晓兰,钟义信. 基于文本理解的自动文摘系统研究与实现[J]. 电子学报,1998,26(7):155 – 158.
YANG Xiao-lan, ZHONG Yi-xin. Study and realization for text interpretation and automatic abstracting[J]. Acta Electronica Sinica, 1998, 26(7): 155 – 158. (in Chinese)

[4] 李维刚,刘挺,李生. 基于网络挖掘的实体关系元组自动获取[J]. 电子学报,2007,35(11):2111 – 2116.
LI Wei-gang, LIU Ting, LI Sheng. Automated entity relation tuple extraction using web mining[J]. Acta Electronica Sinica, 2007, 35(11): 2111 – 2116. (in Chinese)

[5] SOON W M, NG H T, LIm D C Y. A machine learning approach to coreference resolution of noun phrases[J]. Computational Linguistics, 2001, 27(4): 521 – 544.

[6] KEHLER A, APPELT D, TAYLOR L, et al. Competitive self-trained pronoun interpretation [A]. Proceedings of HLT-NAACL [C]. Stroudsburg, USA: ACL, 2004. 33 – 36.

[7] MÜLLER C, RAPP S, STRUBE M. Applying co-training to reference resolution[A]. Proceedings of the 40th Annual Meeting on ACL[C]. Stroudsburg, USA: ACL, 2002. 352 – 359.

[8] RAGHAVAN P, FOSLER-LUSSIER E, LAI A M. Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features[A]. Proceedings of the 2012 Conference of the NAACL[C]. Stroudsburg, USA: ACL, 2012. 731 – 741.

[9] NG V, CARDIE C. Weakly supervised natural language learning without redundant views[A]. Proceedings of the Conference of the NAACL[C]. Stroudsburg, USA: ACL, 2003. 94 – 101.

[10] CHARNIAK E, ELSNER M. EM works for pronoun anaphora resolution[A]. Proceedings of the 12th Conference of the EACL[C]. Stroudsburg, USA: ACL, 2009. 148 – 156.

[11] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples[J]. The Journal of Machine Learning Research, 2006, 7(11): 2399 – 2434.

[12] KULIS B. Metric learning: A survey[J]. Foundations and Trends in Machine Learning, 2012, 5(4): 287 – 364.

[13] STOYANOV V, GILBERT N, CARDIE C, et al. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art[A]. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP[C]. Stroudsburg, USA: ACL, 2009. 656 – 664.

[14] 周炫余,刘娟,邵鹏,等. 基于层次过滤模型的中文指代消解研究[J]. 吉林大学学报(工学版),2016,46(4):1209 – 1215.
ZHOU Xuan-yu, LIU Juan, SHAO Peng, et al. Research of Chinese anaphora resolution based on multi-pass sieve model[J]. Journal of Jilin University (Engineering and Technology Edition), 2016, 46(4): 1209 – 1215. (in Chinese)

[15] CHIANG L H, RUSSELL E L, BRAATZ R D. Fisher Discriminant Analysis//Fault Detection and Diagnosis in Industrial Systems[M]. London: Springer, 2001. 57 – 70.

[16] 孔芳,周国栋. 基于树核函数的中英文代词消解[J]. 软件学报,2012,23(5):1085 – 1099.
KONG Fang, ZHOU Guo-dong. Pronoun resolution in English and Chinese languages based on tree kernel[J]. Journal of Software, 2012, 23(5): 1085 – 1099. (in Chinese)

[17] JAIN P, KULIS B, DHILLON I S. Inductive regularized learning ofkernel functions[A]. Proceedings of the NIPS [C]. Massachusetts, USA: MIT Press, 2010. 946 – 954.

[18] MIKA S, RÄTSCH G, WESTON J, et al. Fisher discriminant analysis with kernels[A]. Proceedings of the 1999 IEEE Signal Processing Society Workshop[C]. Portland, USA: IEEE, 1999. 41 – 48.

[19] MARC V, JOHN B, JOHN A, et al. A model theoretic coreference scoring scheme[A]. Proceedings of the 6th Message Understanding Conference [C]. Stroudsburg, USA: ACL, 1995. 45 – 52.

[20] AMIT B, BRECK B. Algorithms for scoring coreference chains [A]. Proceedings of LREC [C]. Stroudsburg, USA: ACL, 1998. 563 – 566.

[21] LUO X Q. On coreference resolution performance metrics [A]. Proceedings of HLT-EMNLP [C]. Stroudsburg,

USA:ACL,2005. 25 – 32.
[22] NONG Y. The Handbook of Data Mining[M]. Cleveland:CRC Press,2001. 247 – 277.
[23] MARTA R, EDUARD H. BLANC: Implementing the Rand Index for coreference evaluation[J]. Natural Language Engineering,2011,17(4):485 – 510.

作者简介

周炫余 男,1987 年 10 月出生,湖南邵阳人. 现为武汉大学计算机学院博士研究生,主要从事指代消解、中文自然语言处理、机器学习等有关研究.
E-mail:zhouxuanyu@ whu. edu. cn

刘 娟(通讯作者) 女,1970 年 2 月出生,湖北武汉人,教授、博士生导师. 现为武汉大学计算机学院软件所所长,主要从事自然语言处理、生物信息、生物医学图像分析、数据挖掘、机器学习等有关研究.
E-mail:liujuan@ whu. edu. cn