

基于双代价参数 SVM 的生物医学文本指代消解研究

张丽君¹, 李丽双^{*2}, 范国龙²

(1. 辽宁医学院 现代教育技术中心, 辽宁 锦州 121001;

2. 大连理工大学 计算机科学与技术学院, 辽宁 大连 116024)

摘要: 生物医学文本中的指代消解是生物医学信息抽取领域的一个重要组成部分. 通过引入双代价参数对基本 SVM 方法进行改进, 并在 FlyBase 语料集上进行了测试, 准确率、召回率、 F 值分别达到 53.9%、69.5%、60.7%. 同时研究了特征向量的选择和取值对于实验结果的影响. 最后与其他先进方法进行了对比. 结果表明, 在同样的语料上, 基于双代价参数 SVM 方法优于其他先进的方法.

关键词: 生物医学文本; 指代消解; SVM; 双代价参数

中图分类号: TP391

文献标识码: A

doi: 10.7511/dllgxb201504011

0 引言

随着计算机的广泛应用和信息科技的迅猛发展, 生物医学文献呈现井喷式增长. 如何有效地在海量医学文献中挖掘有用的信息, 已成为该领域研究者面临的重要课题. 信息抽取 (information extraction, IE) 是解决这些问题的有效方法. 信息抽取是指从文本中抽取特定的事实信息^[1], 包括命名实体识别、关系抽取以及事件抽取 3 个子任务. 尽管这 3 个任务被视为信息抽取的标准任务, 但为了获得更高的信息抽取性能, 指代消解^[2]越来越受到重视, 被认为是信息抽取任务中不可或缺的一部分.

生物医学领域指代消解方法可以分为 3 类: 基于规则的方法, 基于统计机器学习的方法, 以及规则与统计相结合的方法.

Castano 等^[3]从语义角度制定规则, 在自己标注的 MEDLINE 摘要上进行测试实验, 精确率达到 77%, 召回率达到 71%. Kim 等^[4]利用不同的规则将照应语是代词和名词短语两种情况分开进行消解, 代词采用的是 Grosz 等提出的中心理论^[5], 同时针对名词短语制定了 7 条规则, 在自己标注的 MEDLINE 摘要上进行实验, 取得 59.5%

的精确率和 40.7% 的召回率. Lin 等^[6]利用 UMLS 本体和 SA/AO (subject-action/action-object) 模式从生物语料中挖掘模板进行指代消解. 同样是在自己标注的 MEDLINE 语料上进行实验, 对照应语为代词进行消解的 F 值达到了 92%, 对非代词做照应语消解的 F 值达到了 78%. Miwa 等^[7]单纯使用基于规则的方法进行指代消解, 针对 BioNLP 2011 评测提供的生物医学指代消解语料进行实验. 在开发集上的 F 值达到 60.5%, 在测试集上 F 值为 55.9%, 比 Tuggener 等^[8]在该测试集上的结果提高了 24.94% (55.90% vs. 30.96%).

Torii 等^[9]采用最大熵模型进行分类, 并选取了更适合于生物医学指代消解的特征, 在自己标注的 MEDLINE 摘要上进行实验, 主要针对照应语是限定性的词进行消解, 获得 78% 的准确率. 与不适用这些高性能特征的实验相比, 减少了 10% 的错误率. Gasperin 等^[10]于 2008 年在自己提供的标准的基于全文的生物医学指代消解语料上进行研究. 分类器采用的是贝叶斯模型. 该语料只对名词短语进行消解, 用两种评价方式进行评价, 分别达到 55.5% 和 68.3% 的 F 值.

基于规则的方法简单、易于实现,但需要人工总结规则;机器学习方法,不需人工总结规则模式,鲁棒性好,但需要大规模已标注标准语料;规则和机器学习相结合的方法,对于不同的照应语类型,采用不同的方法,是目前消解性能最好的一种方法,但需要针对不同的情况制定相应的消解方式,操作复杂,普适性差。

基于以上分析,同时结合目前生物医学指代消解方法并不成熟的现状,本文对消解效果较好的支持向量机(support vector machine,SVM)算法进行改进,通过引入双代价参数进行指代消解,并在 FlyBase 语料集上进行测试,以期达到较高的准确率、召回率、 F 值。

1 基于 SVM 分类算法原理

SVM 是 Cortes 和 Vapnik 于 1995 年首次提出的,它在解决小样本、非线性以及高维模式识别中表现出许多特有的优势,并能够推广应用到函数拟合等其他机器学习问题中^[11]。

设原始输入空间 $\mathbf{X} \subseteq \mathbf{R}^n$ (n 为输入空间的维数),训练集 $\mathbf{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$. $\mathbf{x}_i \in \mathbf{X}; y_i \in \{-1, 1\}$, 是 $\mathbf{x}_i$ 的标记,若 $\mathbf{x}_i$ 属于正类, $y_i = 1$, 若 $\mathbf{x}_i$ 属于负类, $y_i = -1$; N 为样本的个数. SVM 就是寻找能够将训练数据划分为两类的最优超平面,该超平面可以通过求下面凸二次规划问题的解得到:

$$\begin{aligned} \max \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i y_i \alpha_j y_j K(\mathbf{x}_i \cdot \mathbf{x}_j) \\ \text{s. t.} \quad & \sum_{i=1}^N y_i \alpha_i = 0; 0 \leq \alpha_i \leq C, i = 1, 2, \dots, N \end{aligned} \tag{1}$$

其中 $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$, 为 Kernel 函数; α_i 为与每个样本对应的 Lagrange 乘子; $C > 0$, 是自定义的惩罚系数. 给定一个测试实例 $\mathbf{x}$, 它的类别由下面的决策函数决定:

$$f(\mathbf{x}) = \text{sgn} \left[\sum_{\mathbf{x}_i \in \mathbf{v}_s} \alpha_i y_i K(\mathbf{x}_i \cdot \mathbf{x}) + b \right] \tag{2}$$

$\mathbf{v}_s$ 为支持向量.

2 基于双代价参数 SVM 的生物医学文本指代消解

2.1 语料的选择和介绍

本文采用的语料,是由 Gasperin 标注的

FlyBase 语料,可以从 FlySlip 官网上获得 (<http://www.sequenceontology.org>). 该语料使用 xml 标注了 5 篇文章,大约有 33 600 个词和 2 629 个被标注了的生物名词短语,这些命名实体类型均基于 sequence ontology^[12],每一个都有 id 的属性,用以唯一标识每一个实体表述. 该语料共标注了两种生物医学实体表述间的关系,即共指关系和联想关系。

在 FlyBase 语料中,还有一个重要的概念,就是生物类型(biotype)的概念. 所有的生物医学命名实体(表述)都根据 biotype 进行了分类,这些实体或者是 gene 类型或者与 gene 存在着某种关系. 表 1 给出了一些常用生物类型的实例。

表 1 生物命名实体的生物类型(biotype)的分布
Tab. 1 The distribution of biotype in creature named entities

生物类型	分布
product	1 154
gene	242
part-of	190
variant	169
otherbio	442
partof_product	239
supertype	36
subtype	156
unsure	1

2.2 语料的预处理

本文对原始语料进行了如下的预处理。

(1)名词短语编号标注. 对每个生物名词短语都增加了 anaph_index 属性,用来标识名词短语的编号. 对于每篇文章,anaph_index 从 1 开始编号,anaph_index 属性的增加主要是刻画距离特征,便于后续的具体特征的提取。

(2)名词短语句子编号标注. 处理的方法与(1)相似,给每个生物名词短语都增加了 sentence_index 属性,用来标识名词短语在文中所在句子的编号,sentence_index 也是从 1 开始编号。

(3)名词短语形式标注. 将名词短语分成以下几类:

- ①限定性名词短语(definite NPs),包括那些以 the 打头的名词短语. 这类名词短语在语料中出现的频率非常高。
- ②非限定性名词短语(indefinite NPs),包括

那些泛指某一特定名词的名词短语,在文本中,将以 a/an/all/both/any/each/every/other 打头的名词短语划为该类. 这类名词短语所描述的对象一般并不确定.

③指示性的名词短语(demonstrative NPs),是那些包含 this/that、these/those、which、whose、its 等指示性代词的名词短语. 指示性的名词短语的先行词和指代词之间的距离一般不会太远.

④量化性的名词短语(quantified NPs),是那些包含诸如 one、some、several、at least、many 等数量词的名词短语.

⑤专用名词短语(proper NPs). 将剩下的那些包含基因的名词短语称为专用名词短语. 这里专用名词短语的定义与通常的不同,这是由于本语料的名词短语或多或少与基因存在着某种关系.

⑥未归类的名词短语(other NPs). 将余下的名词短语统一划归为 other NPs.

根据上述分类,给每个名词短语增加一个 form 属性,属性值为 def、indef、dem、quant、pn、other 之一.

(4)最近含有相同基因的 id 标注. 由于本语料中的名词短语多数含有基因名,含有相同基因并且离得越近越有可能构成共指关系. 所以,增加了 nearest_samegen_id 属性,属性值为满足上述条件的 id.

2.3 基于双代价参数 SVM 的生物医学文本指代消解模型

本文采用改进后的双代价参数 SVM 模型进行指代消解.

基本 SVM 模型中所选择的代价函数为

$$\begin{aligned} \min J(\mathbf{w}, b, \boldsymbol{\xi}) &= \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \\ \text{s. t. } y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) &\geq 1 - \xi_i; \\ \xi_i &\geq 0, i = 1, 2, \dots, N \end{aligned} \tag{3}$$

其中 C 是惩罚因子.

C 表征了对离群点的重视程度, C 越大,表明越重视离群点. 实际过程中, C 取得越大,训练样例的正确分类程度越高. 但是由于实际正负训练样例相当不平衡,需要解决分类问题中样本偏斜问题.

为了解决数据集的偏斜问题,本文将赋予数量少的正训练样例更大的惩罚因子(代价),表明更加重视这部分样本,这里采用双代价参数的方法,因此目标函数变为

$$\min J(\mathbf{w}, b, \boldsymbol{\xi}) = \frac{1}{2} \|\mathbf{w}\|^2 + C_+ \sum_{i=1}^p \xi_i + C_- \sum_{i=p+1}^N \xi_i \tag{4}$$

其中 $i = 1, \dots, p$ 都是正的训练样例; $i = p + 1, \dots, N$ 都是负的训练样例.

2.4 特征的选择

选择有意义的特征是进行基于机器学习的指代消解的前提,本文通过分析生物名词特点和前人研究的经验选择了 15 个特征. 采用递增式学习策略^[13],首先初始化最小完备特征集 F^* 为原子特征集(包括照应语的形式、候选先行语的形式、单复数一致性 3 个基本的词法特征),然后每次加入一条特征项,若识别效果提高则将该特征项加入特征集 F^* 中,否则放弃该特征项. 如此循环,直到添加完所有特征项,最后得到的特征集 F^* 就是最小完备的特征集. 最小完备的特征集共包括以下 11 个特征:

(1)词法特征

①照应语的形式(Fa)

名词短语的形式表明了名词短语的使用方式,照应语的形式有助于表达名词短语的类型,例如,不定描述的名词短语表明所描述的对象不确定,指示性的名词短语的先行词和照应词之间的距离一般不会太远.

②候选先行语的形式(Fc)

这个特征刻画候选先行语的形式,特征的取值与 Fa 的值一致.

③单复数一致性(Num)

构成共指关系的名词短语通常单复数一致,因此单复数一致的名词短语更有可能构成共指关系,这个特征正好刻画照应语与候选先行语之间单复数是否一致.

④修饰词匹配(Mm)

这个特征刻画了构成共指关系的候选先行语和照应语的修饰词特征,具有共指关系的候选先行语和照应语的修饰词更有可能匹配. 这个特征显示了候选先行语和照应语是否含有至少一个相同的修饰词,如果存在,就说明候选先行语和照应

语的修饰词匹配.

(2)距离特征

距离特征是候选先行语和照应语之间距离的度量.与照应语跨度越小的先行词越有可能构成共指关系.研究者使用句子跨度和实体跨度两种距离特征在不同层次上对距离进行度量.

①句子跨度指标(Sd)

这个特征通过计算候选先行语和照应语句子编号的跨度进行度量.假定候选先行语与照应语的句子跨度越小,越有可能构成共指关系.

②实体跨度指标(Ed)

这个特征通过计算候选先行语和照应语实体编号的跨度进行度量.同样,假定候选先行语与照应语的实体跨度越小,越有可能构成共指关系.

(3)领域知识特征

为了深度挖掘名词短语之间的关系,生物学中的指代消解需要利用具体的领域知识,提取相关特征.文本共提取了 5 个领域知识方面的特征.

①照应语的生物类型(Ta)

每一个生物命名实体都被划分为一类语义类型.在本算法中,生物命名实体被分为 9 类生物类型,即 product、gene、partof、variant、otherbio、partof_product、supertype、subtype、unsure.本语料中共有 2 629 个命名实体.候选先行语和照应语的生物类型具有某种关联,具有共指关系的先行语和照应语更有可能具有相同的生物类型.

②候选先行语的生物类型(Tc)

这个特征刻画候选先行语的生物类型,特征值与 Ta 的值一致.

③生物类型一致性(Sa)

这个特征描述了候选先行语和照应语的生物类型是否相同.假定生物类型越相同的名词短语越有可能构成共指关系.

④含有基因(Gn)

语料中描述生物命名实体的名词短语,很多都包含基因名(例如 neur activity 是 otherbio 类型,但是 neur 是一个基因名).本语料 70.8%的生物命名实体含有基因名,其中构成共指关系的先行语-照应语对中 69.9%都含有基因名,并且 88.7%含有相同的基因名,仅 11.3%含有不同的基因名.

因此该特征会极大地影响指代消解系统的性能.根据候选先行语和照应语是否含有基因以及基因是否相同将它分为 5 类,即候选先行语和照应语都不含有基因;候选先行语含有基因,照应语不含有基因;候选先行语不含有基因,照应语含有基因;候选先行语和照应语都含有基因,且相同;候选先行语和照应语都含有基因,且不同.

⑤构成邻近基因(Sgn)

该特征将那些候选先行语和照应语含有相同的基因细化,将它们分成两类:距离最近的划为一类,剩下的划为一类.

3 实验与结果

3.1 基本 SVM 的测试结果

本文采用 10 倍交叉验证法,将训练样例处理为相应的格式,作为 SVM 的输入.使用支持向量机 LibSVM 系统,代价参数(惩罚因子) $C=100$ 时已经稳定,核函数选用径向基函数,结果如表 2 所示.

表 2 FlyBase 语料上的 10 倍交叉验证结果
Tab. 2 10-Fold cross-validation results of FlyBase corpus

实验次数	准确率/%	召回率/%	F 值/%
1	67.1	62.5	64.7
2	70.4	50.6	58.9
3	85.9	38.1	52.8
4	56.0	37.5	44.9
5	60.7	40.6	48.6
6	84.2	30.0	44.2
7	83.0	30.6	44.7
8	65.5	50.0	56.7
9	81.8	61.8	70.4
10	89.7	44.3	59.3
平均值	74.4	44.6	55.8

从表 2 可以看出,测试结果获得的平均性能是准确率 74.4%、召回率 44.6%、F 值 55.8%.

3.2 双代价参数 SVM 的测试结果

采用 2.3 节所建立的双代价参数 SVM 模型进行改进实验.因为实验数据中负例样本远远多于正例样本,所以设置双代价参数如下:

固定 $C = C_+$, 令

$$\epsilon = C_- / C_+ \tag{5}$$

其中 $0 < \epsilon \leq 1$,通过参数调优的方式来调节 ϵ ,从而获得相对最优的效果.

ϵ 对平均性能各指标 P 的影响如图 1 所示。

图 1 中表明了 ϵ 的取值对召回率、准确率、 F 值平均性能的影响。从图中可以看出, ϵ 越大, 准确率越高, 召回率越低, 而 F 值在 $\epsilon = 0.37$ 时取得最高值, 即为 60.7%。

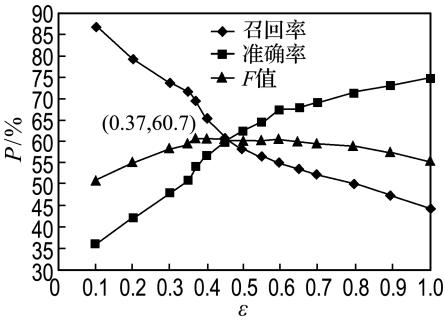


图 1 ϵ 对平均性能各指标的影响

Fig. 1 The influence of ϵ on the index of average performance

$C = C_+ = 100, \epsilon = 0.37$ 时做的 10 次实验结果如表 3 所示。

表 3 $\epsilon = 0.37$ 时 10 倍交叉验证结果

Tab. 3 10-Fold cross-validation results when $\epsilon = 0.37$

实验次数	准确率/%	召回率/%	F 值/%
1	53.6	78.1	63.6
2	44.7	72.5	55.3
3	71.8	57.5	63.8
4	39.4	59.3	47.3
5	41.2	69.3	51.7
6	54.9	70.0	61.5
7	52.8	57.5	55.0
8	41.6	71.8	52.7
9	63.3	83.1	71.8
10	75.9	75.9	75.9
平均值	53.9	69.5	60.7

通过表 3 可以看出:使用双代价参数的 SVM 模型,测试结果获得的平均性能是准确率53.9%、召回率 69.5%、 F 值 60.7%。准确率和召回率仅仅相差 15.6%,相比基本 SVM 模型准确率和召回率相差 29.8%得到了较大改进,平衡了召回率和准确率, F 值也得到提高。

3.3 与其他先进方法的指标对比

上述测试结果均采用 FlyBase 语料,目前使用 FlyBase 语料的只有 Gasperin 基于概率决策理论建立的模型(称为 Prob+Closest 模型)进行

的测试,表 4 将基于同一语料的 3 种模型的测试关键指标进行了对比。

表 4 3 种模型的 3 个指标对比

Tab. 4 The comparison of three indicators in three models

模型	准确率/%	召回率/%	F 值/%
Prob+Closest	66.2	50.0	56.9
基本 SVM 模型	74.4	44.6	55.8
双代价参数 SVM 模型	53.9	69.5	60.7

在表 4 中,第一种是 Gasperin 使用的 Prob+Closest 模型;第二种是使用基本 SVM 建立的模型;第三种是使用双代价参数建立的模型。双代价参数 SVM 模型获得的 F 值最高,比 Gasperin 的 Prob+Closest 模型的 F 值提高了 3.8%。

4 结 语

本文首先针对特征的选择和特征取值的定义方法进行了具体的分析和描述。然后通过引入双代价参数对传统 SVM 算法进行了改进,在 FlyBase 语料上进行测试,最终获得的准确率、召回率、 F 值分别为 53.9%、69.5%、60.7%。双代价参数 SVM 模型比 Gasperin 的 Prob+Closest 模型的 F 值提高了 3.8%,说明使用双代价参数对基本 SVM 模型进行改进明显优于其他先进方法。

如何采用某种算法(或者规则限定),获取更加平衡的正负训练样例,以及如何采取有效措施,克服 SVM 对于正负训练样例不平衡导致的偏斜现象还需要进一步深入探索。

参考文献:

[1] 李保利,陈玉忠,俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 39(10):1-5, 66.
LI Bao-li, CHEN Yu-zhong, YU Shi-wen. Research on information extraction: A Survey [J]. Computer Engineering and Applications. 2003, 39(10):1-5, 66. (in Chinese)

[2] Ng V. Supervised noun phrase coreference research; the first fifteen years [C] // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL). Uppsala: Cambridge University Press, 2010:1396-1411.

- [3] Castano J, Zhang J, Pustejovsky J. Anaphora resolution in biomedical literature [C] // **Proceedings of the 2002 International Symposium on Reference Resolution**. Alicante:[s n], 2002:1-6.
- [4] Kim J, Jong C P. BioAR: anaphora resolution for relating protein names to proteome database entries [C] // **Proceedings of the Reference Resolution and Its Application Workshop in Conjunction with ACL**. Barcelona:ACL, 2004:79-86.
- [5] Grosz B J, Joshi A K, Weinstein S. Centering: a framework for modeling the local coherence of discourse [J]. **Computational Linguistics**, 1995, **21**(2):203-225.
- [6] Lin Y H, Liang T. Pronominal and sortal anaphora resolution for biomedical literature [C] // **Proceedings of the 16th Conference on Computational Linguistics and Speech Processing**. Taipei: Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 2004: 101-110.
- [7] Miwa M, Thompson P, Ananiadou S. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution [J]. **Bioinformatics**, 2012, **28**(13):1759-1765.
- [8] Tuggener D, Klenner M, Schneider G, *et al.* An incremental model for the coreference resolution task of BioNLP 2011 [C] // **Proceedings of BioNLP Shared Task 2011 Workshop**. Portland: Association for Computational Linguistics, 2011:151-152.
- [9] Torii M, Vijay-Shanker K. Anaphora resolution of demonstrative noun phrases in MEDLINE abstracts [C] // **Proceedings of 2005 Pacific Asia Conference on Computational Linguistics**. [s l]: Cambridge University Press, 2005:332-339.
- [10] Gasperin C, Briscoe T. Statistical anaphora resolution in biomedical texts [C] // **Coling 2008 - 22nd International Conference on Computational Linguistics, Proceedings of the Conference**. East Stroudsburg: Cambridge University Press, 2008: 257-264.
- [11] Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers [C] // **Proceedings of 5th Annual ACM Workshop on Computational Learning Theory**. New York:ACM, 1992:144-152.
- [12] Eilbeck K, Lewis S E. Sequence ontology annotation guide [J]. **Comparative and Functional Genomics**, 2004, **5**(8):642-647.
- [13] 李丽双, 党延忠, 廖文平, 等. CRF 与规则相结合的中文地名识别 [J]. 大连理工大学学报, 2012, **52**(2):285-289.
LI Li-shuang, DANG Yan-zhong, LIAO Wen-ping, *et al.* Recognition of Chinese location names based on CRF and rules [J]. **Journal of Dalian University of Technology**, 2012, **52**(2):285-289. (in Chinese)

Research on anaphora resolution in biomedical texts based on doubly-cost parameter SVM

ZHANG Li-jun¹, LI Li-shuang^{*2}, FAN Guo-long²

(1. Center for Educational Technology, Liaoning Medical University, Jinzhou 121001, China;

2. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: The anaphora resolution in biomedical text is an important part of biomedical information extraction. The basic SVM method was improved by introducing doubly-cost parameters, and tests on the FlyBase corpus were conducted. The precision, recall and *F*-value are 53.9%, 69.5% and 60.7% respectively. In the meantime, the influence on the final results of different selections of the feature vectors are studied. Finally, compared with other state-of-the-art methods on the same corpus, it is shown that the proposed SVM method based on doubly-cost parameters is superior to the others.

Key words: biomedical texts; anaphora resolution; SVM; doubly-cost parameters