自然语言文本中不确定性信息的自动识别①

杨文敏, 李保利

(河南工业大学 信息科学与工程学院, 郑州 450001)

摘 要: 自然语言中存在大量不确定的表述, 针对此类信息的检测任务是信息抽取领域的研究热点之一, 然而, 面向中文的不确定信息检测研究仍然比较匮乏, 利用支持向量机(Support Vector Machine, SVM)能够很好的解决非线性、高维数、局部小样本等实际问题的优势, 将中文不确定性信息识别问题转化为分类问题, 通过在复旦大学发布的中文不确定性检测数据集语料上的实验, 验证了本文提出的基于 SVM 的中文不确性信息检测方法的有效性, 相比于句子评分模型, 我们的系统取得了更好的召回率.

关键词: 不确定性信息检测; 支持向量机; 语料; 分类

Automatic Identification of Uncertainty Information in Natural Language

YANG Wen-Min, LI Bao-Li

(Henan University of Technology, College of Information Science and Engineering, Zhengzhou 450001, China)

Abstract: There are a lot of uncertainty information in natural Language. It is becoming a focus of researches in NLP recently. However, the research on Chinese is still scarce. In this paper, we use SVM, which is a good solution to the high dimension, nonlinear and local small sample, to identify Chinese uncertainty information recognition as a question of classification. We carry experiment on Chinese uncertainty corpus published by Fudan University, which confirms the availability exposed by our paper based on SVM. Compared to the sentence scoring model, our system has better recall rate.

Key words: uncertainty information detection; SVM; corpus; classification

1 引言

信息抽取是直接从自然语言文本中抽取事实信息, 然而在自然语言中常常混杂有猜测性的、不确定性的信息. 如果不加甄别或者不能准确识别一个自然语言句子表达的是不确定性的信息还是确定无疑的事实, 那么由于抽取出的信息的不确定性而导致的误报率的上升, 必然会使信息抽取系统的准确率大大降低. 然而在现实的信息抽取系统中都是假设抽取的信息是准确的, 可是从言语使用的现实情况看来, 这样的假设并不是总成立的. 假如我们可以将确定的信息和不确定的信息分开, 就可以分别做处理. 因此, 如何有效的识别自然语言文本中不确定性信息就成为提高信息抽取系统准确率与效率的关键技术.

汉语言学者对不确定信息作了如下定义: 不确定信息指事物类属边界或性质状态不明确, 是对事物性质状态模糊认识的反映[1].

例句 1: 明天可能会下雨.

例句2: 如果发现问题比较严重和复杂, 时间会相应延长.

例句 3: 珠穆朗玛峰是世界最高峰.

例 1 中推测明天可能会下雨, 至于明天会不会下雨是不确定的. 例 2 中假设“发现问题比较严重和复杂”结果会使时间相应延长, 至于事实是不是这样是不确定的. 例 3 陈述了一个事实, 不存在不确定的信息.

2 相关工作

虽然这方面的相关研究不多, 可是从语言学角度

① 基金项目:河南省基础与前沿技术研究项目(112300410007)
收稿时间:2014-05-20;收到修改稿时间:2014-11-03

来看, 语言中隐含的不确定性很早就有语言学家关注, 从机器学习角度研究如何识别自然语言文本中的不确定信息近些年才受到关注. 早期的研究都是通过手工编制的关键词表来识别文章中的猜测性的句子, 这些方法有明显的缺点, 不仅需要大量的人力来总结关键词词典, 正确率不高同时也受到专家领域知识的约束. Light 等[2]人在 2004 年使用了一个手工编制的关键词表, 利用匹配的方法来识别医学文章中的猜测性的句子. 该方法取得了较高的召回率(80%), 然而精确率却很低(50%).

由于我们很难用有限的规则来准确刻画出含有不确定信息的句子, 越来越多的研究人员开始借助机器学习算法来进行不确定信息检测的研究. Medlock 与 Briscoe(2007)[3]使用词本身作为特征, 利用弱监督的方法将生物学文章的句子分为猜测性的和非猜测性的. Kilicoglu 和 Bergler[4]在 2008 年提出使用句法信息以便半自动的精化不确定信息词表.

Vincze 等人于 2008 年发布了 BioScope 语料和 Wikipedia 语料, 构建了英文生物学文献不确定线索词及其所限制的范围的语料库, 为训练、测试和比较生物医学文献中不确定信息监测提供了公共的英文语料资源, 成为从事此项研究的一个重要基础资源. Morante 和 Daelemans(2009)[5]等借助此语料, 采用决策树算法来识别句子中是否包含有不确定信息.

此外, 目前的研究多以英文文本为处理对象, 中文不确定性信息检测很少. 复旦大学计峰等[6]人在 2010 年利用不确定性句子中普遍存在的线索词的信息构建了句子的评分模型, 用来识别不确定信息.

本文从句子层面识别不确定信息,提出了基于 SVM 的中文不确性信息识别模型, 判断句子中是否含有不确定信息. 实验表明, 在不确定信息检测任务中, 基于 SVM 的中文不确定性信息识别模型可以获得较好的性能.

3 支持向量机模型

支持向量机(Support Vector Machine)是 Cortes 和 Vapnik 于 1995 年首先提出的, 该算法是从线性可分情况下的最优分类面发展而来的, 它的主要思想可用图 1 的两维情况说明. 假设训练样本(x_1, y_1), …, (x_m, y_m), $x \in R^n$, $y \in \{+1,-1\}$, m 为样本数, 分类问题可以描述为:

$$y_i(w \cdot x_i + b) - 1 \geq 0 . \quad (1)$$

其中, x_i 是输入的第 i 个样本; $y_i \in \{+1,-1\}$ 表示相应 x_i 期望分类值; w 是可调权值向量; b 是偏置. 分类间隔为:

$$\gamma = \min_{x_i|y_i=1} \frac{w \cdot x_i + b}{\| w \|_2} - \min_{x_i|y_i=-1} \frac{w \cdot x_i + b}{\| w \|_2} = \frac{2}{\| w \|_2} \quad (2)$$

分类问题的求解就是在条件式(1)下最大化式(2), 这是一个典型的二次规划问题, 因此目标函数为

$$\min_w L(w,b,\alpha) = \frac{1}{2} \| w \|_2^2 - \sum_{i=1}^{m} \alpha_i [y_i (w \cdot x_i + b) - 1] \quad (3)$$

其中, $\alpha_i \geq 0$ 是 lagrange 乘子.

此优化问题可转化为求解其“对偶”问题, 即:

$$\max_\alpha W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i, x_j) \quad (4)$$

约束条件: $\sum_{i=1}^{m} \alpha_i y_i = 0$ (5)

对于非线性问题, 目标函数变为:

$$\max_\alpha W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (6)$$

最后训练后相应的分类函数(即支持向量机)为:

$$d(x) = \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b^* \quad (7)$$

通过 d(x)的符号来确定样本 x 的分类情况.

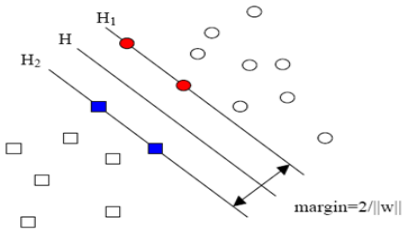

图 1　最优分类面

圆和方分别代表两个不同的类, H 为最优的分类面. H1、H2 之间的距离 margin 就是两类之间的分类间隔.

4 实验及分析

4.1 文本的预处理

预处理阶段包括中文分词、文本表示和特征提取 3 个步骤. 首先用中科院分词系统 ICTCLAS 对文本进行分词, 然后用 TF-IDF 进行权重计算, 最后将文档

用向量空间模型(vector space model, VSM)表示. 经典的 TF-IDF 计算公式如下:

$$w_{ij}=tf_{ij}\times idf_j=tf_{ij}\times\log(\frac{N}{n_j}) \tag{8}$$

其中 tf_{ij} 指特征项 t_j 在文档 d_i 中出现的次数; idf_j 表示出现特征项 t_j 的文档数的倒数. N 表示总文本数, n_j 表示出现特征项 t_j 的文档数.

在很多情况下还需要将向量归一化, TF-IDF 的归一化计算公式如下:

$$w_{ij}=tf_{ij}\times idf_i\times wa_{ij}=\frac{tf_{ij}\times\lg(\frac{N}{df_i}+\alpha)}{\sqrt{\sum_{k_i\in d_j}[tf_{ij}\times\lg(\frac{N}{df_i}+\alpha)]^2}} \tag{9}$$

其中, 分母为归一化因子, α 为经验常数, 一般取 0.01.

4.2 语料收集与数据初步统计

由于缺乏中文的标注语料, 我们使用复旦大学发布的中文不确定性检测数据集作为实验语料, 该语料选取常见的 150 个线索词作为查询词, 例如表假设的词“如果”, “可能”等; 表将来时的词“将”, “预计”等; 表示个人信仰的词“相信”“认为”等. 总共 10000 个句子, 选取其中 8000 句作为训练语料, 2000 句作为测试语料. 该语料的初步统计见表 1.

表 1 中文语料统计

	训练集	测试集
句子数量	8000	2000
词数量	245723	61584
不确定性句子数量	2248	610
不确定性句子百分比(%)	28.1	30.5
线索词数量	3982	1102
平均线索词数量	1.77	1.81

4.3 评价指标

按照 CoNLL2010 的评价标准, 本文研究中采用精确率 P、召回率 R 和综合评价指标 F1 值对实验结果进行评价.

$$P=\frac{\text{系统标注正确的句子数}}{\text{系统标出的句子总数}}*100\% \tag{10}$$

$$R=\frac{\text{系统标注正确的句子数}}{\text{测试集中出现的句子总数}}*100\% \tag{11}$$

$$F_1=\frac{2\times P\times R}{P+R}\times 100\% \tag{12}$$

4.4 实验结果分析

在模型训练时, 我们按照上述步骤进行预处理, 得到特征向量空间. 选取 Libsvm-3.17 进行模型的训练, 选择合适的参数构造分类器, 直到分类器具有较好的分类能力. 训练结果如表 2 所示.

表 2 训练结果统计

	P	R	F
基准系统	0.625	0.7073	0.6636
本文系统	0.6667	0.7269	0.6953

在测试阶段, 用构造好的分类器去测试测试语料, 采用十字交叉验证的方法, 以检验我们模型的性能. 测试结果如表 3 所示.

作为对比系统, 我们将“句子评分模型”作为基准系统. 实验过程如图 2 所示.

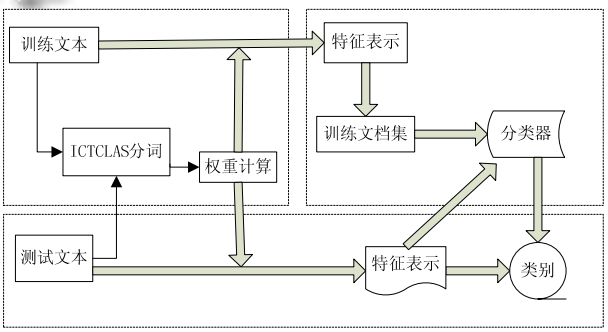

图 2 训练测试过程

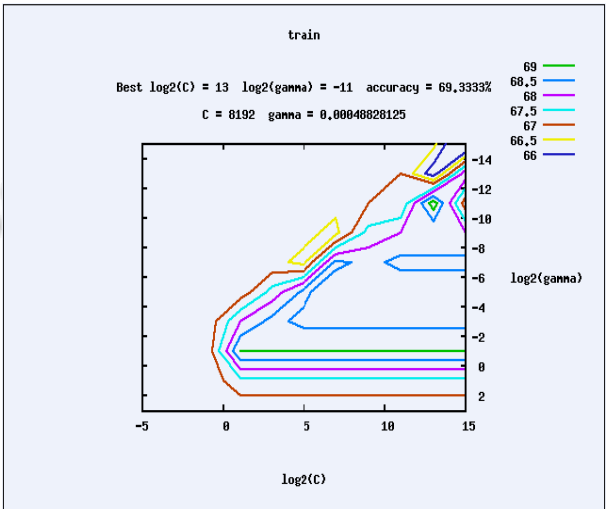

图 3 参数选择过程

训练过程参数选择过程:

参数 c 和 g 最终取值 8192 和 0.00048828125 精确率达到 69.3333%.

训练过程实验结果如表 2 所示.

测试过程实验结果如表 3.

表 3 测试结果统计

	P	R	F
基准系统	0.7024	0.7082	0.7053
本文系统	0.7872	0.8431	0.8142

从实验结果看, 我们的模型虽然在训练集上只取得 0.6953 的 F1 值, 但在测试集上取得了 0.8142 的 F1 值, 同时召回率也相应的提高.

4.5 **错误分析**:

① 像这种情况也出现过两三次. 人工标注结果: 不确定. 测试结果为: 确定

② 沙赫扎德接受过巴基斯坦塔利班一定程度上的培训. 人工标注结果: 不确定. 测试结果为: 确定

我们人工标注时“两三”都是标记为不确定的, 然而分词时将“两三”分为了“两”“三”, 而“两”“三”, 本身是确定的. 同样的还有“一定”都是标记为确定的, 而“一定程度”“一定量的”都是标记为不确定的, 由于分词的时候总是将“一定程度”分为“一定”和“程度”两个词, 所以测试的时候遇到包含“一定程度”、“一定量的”句子都标记为确定的.

5 总结与展望

在本文中, 我们研究了中文不确定信息识别的问题, 实验中使用复旦大学给出的中文不确定性检测数据集语料, 以词作为特征, 提出基于支持向量机模型的中文不确定性句子识别系统, 要求识别出给定文章中哪些句子是不确定的哪些句子是确定的, 经实验证明, 这种方法在中文不确定性句子识别任务中可以取得较高的性能, 通过分析测试错误的句子, 我们发现大部分都是中文的词语出现在不同的句子中表达不同的意思造成的标记错误.

在未来的工作中, 我们可以从以下几方面改进. 第一、我们的实验语料相对较少, 线索词的分类也不够系统, 我们希望在以后的工作中收集更多的语料, 建造一个系统的语料库. 第二, 在特征方面, 本文中我们仅使用了词作为特征, 以后可以考虑用更多的特征, 比如句法特征、标点符号等. 第三、在模型方面, 应该结合更多的中文语言特点, 建立更加适应中文的不确定性句子识别模型. 第四、本文中我们仅在句子级别识别不确定信息, 在以后的工作中, 我们可以考虑从细粒度识别不确定信息, 识别出一句话中到底哪部分是不确定的, 更精确的缩小不确定信息表达的范围.

参考文献

1 黎千驹,模糊语言及相关术语界说.平顶山学院学报, 2008,23(6).

2 Marc L, Qiu XY, Pandmini S. The language of bioscience: facts, speculations, and statements in between. Proc. of the BioLINK, Association for Computational Linguistics. 2004. 17–24.

3 Medlock B, Briscore T. Weakly supervised learning for hedge classification in scientific literature. Proc. of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics. 2007. 992–999.

4 Kilicoglu H, Bergler S. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. Proc. of the Workshop on Current Trends in Biomedical Natural 1Language Processing. Columbus, Ohio. Association for Computational Linguistics. 2008, 6. 46–53.

5 Morante R, Daelemans W. Learning the scope of hedge cues in biomedical texts. Proc. of the Workshop on BioNLP. ACL. 2009. 28–36.

6 计峰,邱锡鹏,黄萱菁.中文不确定性句子的识别研究.第六届全国信息检索学术会议论文集,2010.

7 应伟,王正欧,安金龙.一种基于改进的支持向量机的多类文本分类方法.计算机工程,2006,32(16).

8 李晓艳.生物医学文献中模糊限制语及其范围的测检[学位论文].大连:大连理工大学,2011.

9 Farkas R, Vincze V, Mora G, et al. Cross-genre and cross-domain detection of semantic uncertainty. Proc. of ACL-12, the 50th Annual Meeting of the Association of Computational Linguistics. 2012. 335–367.