基于特征选择的实体关系抽取 *

毛小丽, 何中市, 邢欣来, 刘 莉
(重庆大学 计算机学院, 重庆 400044)

摘 要: 提出了一种实体关系抽取方案,该方案针对实体关系抽取中特征空间维数过高问题,引入了文本分类中的特征选择算法,如信息增益、期望交叉熵和 x^2 统计,实现了特征空间降维。实验结果表明,各特征选择算法均能在尽量保证抽取性能的同时有效地降低向量空间维数,提高分类效率,其中 x^2 统计取得的效果最好。

关键词: 关系抽取; 特征选择; 信息增益; 期望交叉熵; x^2 统计
中图分类号: TP309 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0530-03
doi:10.3969/j.issn.1001-3695.2012.02.034

Entity relation extraction based on feature selection

MAO Xiao-li, HE Zhong-shi, XING Xin-lai, LIU Li
(*College of Computer, Chongqing University, Chongqing* 400044, *China*)

Abstract: This paper proposed a new entity relation extraction method using the feature selection algorithms such as information gain, expected cross entropy, x^2 statistic which were used in text classification for reducing the feature space dimension. Experiments results show that these feature selection algorithms can keep the extraction performance while ensuring reduce the vector space dimension effectively, and improve the classification efficiency. The x^2 statistic reaches the best performance.
Key words: relation extraction; feature selection; information gain; expected cross entropy; x^2 statistic

实体关系抽取是信息抽取研究中的一个重要环节[1],它的作用是抽取出两个实体之间的语义关系。目前,实体关系抽取一般都只考虑一个句子中两个实体间的关系,即实体关系抽取问题的输入是一个句子和句子中已经标记出的两个实体,输出则是这两个实体间的关系。

目前解决实体关系抽取问题主要采用指导性机器学习方法,主流的指导性学习方法有基于特征向量的方法和基于核函数的方法。其中,基于特征向量的方法[2,3]是将关系样例进行特征抽取并将其表示为特征向量,然后通过机器学习的方法来训练关系实例。而基于核函数的方法[4,5]直接以结构树为处理对象来计算它们之间的相似度,再使用支持核函数的分类器进行关系抽取。然而,基于核函数方法的一个致命的缺点是训练和预测的速度太慢,不适于处理大量的数据。因此,本文以基于特征向量的方法解决实体关系抽取问题。

由于基于特征向量的实体关系抽取方法中,特征空间维数一般能达到几万或者几十万维,这样的高维向量一方面将使得训练分类模型以及预测结果的时间开销大大提高,另一方面还可能由于引入了一些不必要的特征而使得抽取性能有所降低。因此,本文考虑将文本分类中的特征选择算法引入到实体关系抽取中,希望能在降低时间开销的同时提高抽取性能。

1 实体关系抽取

1.1 特征抽取

在基于特征向量的实体关系抽取方法中,其首要问题在于有效特征的选择[6]。本文提出的实体关系抽取方案中选择的特征分为五类:实体及其上下文词法特征、动词特征、距离特征、实体扩展特征、语义角色特征[7]。

实体及其上下文特征主要包含实体中心词、实体前两个词、实体后两个词以及这些词的词干和词性。实体及其上下文特征是最基本、最简单的特征。

动词特征表示句子中的所有动词。

距离特征是指要抽取实体关系的两个实体间的词距。

实体扩展特征是指实体的同义词和上位词。

语义角色特征是指用实体的语义角色作为特征。

表 1 是句子"A misty 〈e1〉ridge〈/e1〉 uprises from the 〈e2〉surge〈/e2〉."抽取出的特征,其中实体为"ridge"和"surge"。

表 1 特征文件抽取实例

特征类型	特征
实体及其上下文特征	a_word -2, a_stem -2, ZZ0-POS -2, misty_word -1, misty_stem -1, AJ0-POS -1, ridge_word, ridge_stem, NN1-POS, uprises_word +1, uprise_stem +1, VVZ-POS +1, from_word +2, from_stem +2, PRP-POS +2, from_word -2, from_stem -2, PRP-POS -2, the_word -1, the_stem -1, AT0-POS -1, surge_word, surge_stem, NN1-POS
动词特征	uprise
距离特征	3
实体扩展特征	agent-underspecified, surfaceRegion-underspecified, landmark-underspecified, region-underspecified, ……
语义角色特征	A0 * A2, A0 * A2 * uprise *, uprise from the

1.2 特征选择

通过上面特征抽取的描述可以发现,实体关系抽取问题与

收稿日期: 2011-07-01; **修回日期**: 2011-08-05 **基金项目**: 中央高校基本科研业务费科研专项资助项目(CDJXS11180020);国家科技重大专项项目(2008ZX07315-001)

作者简介: 毛小丽(1986-),女,四川泸州人,硕士研究生,主要研究方向为自然语言处理、机器学习(maoxiaoli0303@163.com);何中市(1965-),男,教授,博导,博士,主要研究方向为机器学习、数据挖掘、自然语言处理;邢欣来(1984-),男,博士研究生,主要研究方向为自然语言处理、复杂网络;刘莉(1986-),女,硕士研究生,主要研究方向为自然语言处理.

文本分类[8,9]问题有相似之处,它们都是采用一串字符作为特征,因此从语料中抽取出的所有特征就组成了原始的特征空间。然而,一个小规模的语料库就要抽取出上万个不同的特征,对于分类器来说,这样的高维空间时间开销是非常大的[10]。因此,希望寻找一种特征选择方法,能在保证分类性能的同时降低空间维数,提高分类效率。对于文本分类问题,已经有很多成熟的特征选择算法用于特征降维。而对于实体关系抽取问题,却还没有相关研究。本文考虑到实体关系抽取问题与文本分类问题的相似性,拟将文本分类中的特征选择算法引用到实体关系抽取中,用于解决实体关系抽取问题中空间维数过高带来的问题。下面对引入的特征选择算法进行介绍。

1.2.1　信息增益

信息增益(information gain,IG)这个概念也是来源于信息论。在实体关系抽取中,它表示了某个特征存在与否对实体关系分类的影响。它的值越大,代表影响越大,因此在用它进行特征选择时,总是选择信息增益大的若干个特征。信息增益的计算公式为

$$IG(t) = P(t)\sum_{j=1}^{m} P(C_j|t)\log(P(C_j|t)/P(C_j)) + P(\bar{t})\sum_{j=1}^{m} P(C_j|\bar{t})\log(P(C_j|\bar{t})/P(C_j)) \quad (1)$$

其中:m 代表实体关系的总类别数;$P(C_j)$ 表示类别为 C_j 的训练句子在整个语料库中出现的概率;$P(t)$ 表示整个语料库中抽取出的特征集合包含特征 t 的概率;$P(\bar{t})$ 表示整个语料库中抽取出的特征集合不包含特征 t 的概率;$P(C_j|t)$ 表示训练句子抽取出的特征集合包含特征 t 时属于 C_j 类的条件概率;$P(C_j|\bar{t})$ 表示训练句子抽取出的特征集合不包含特征 t 时属于 C_j 类的条件概率。

1.2.2　期望交叉熵

期望交叉熵(expected cross entropy, CE)与信息增益的区别在于:信息增益考虑了一个特征在训练句子中存在和不存在两种情况,而期望交叉熵只考虑了特征在训练句子中存在的情况。它的计算公式为

$$CE(t) = P(t)\sum_{j=1}^{m} P(C_j|t)\log(P(C_j|t)/P(C_j)) \quad (2)$$

1.2.3　χ^2 统计(CHI)

χ^2 统计在统计学中是用于度量两个变量之间的相关性的。在实体关系抽取中,用于度量特征与类别之间的相关程度,在这里假设特征与类别之间符合具有一阶自由度的 χ^2 分布。在实际应用中,采用其近似公式为

$$\chi^2(C_j,t) = (AD-CB)^2 \times (A+B+C+D)/((A+C)\times(B+D)\times(A+B)\times(C+D)) \quad (3)$$

其中:A 表示属于 C_j 类并包含特征 t 的训练句子频率;B 表示不属于 C_j 类但包含特征 t 的训练句子频率;C 表示属于 C_j 类但不包含特征 t 的训练句子频率;D 表示不属于 C_j 类也包含特征 t 的训练句子频率。为了得到一个特征对实体关系抽取的重要程度,可以将 $\chi^2(C_j,t)$ 进行加权求和,和值越大代表该特征对实体关系抽取越重要。

1.3　实体关系抽取方案

按照本文提出的实体关系抽取方案,根据上面描述的一系列步骤,利用 SVM 算法构造分类器以判断实体关系类型。本文使用的实体关系抽取方案具体步骤如下:

a)原始语料预处理。对原始语料进行词性标注、句法分析和语义角色标注。

b)特征抽取及特征向量构造。对于语料里每条句子中的实体对,先从预处理后的文本中抽取出上文描述的特征,然后将抽取出的每个特征值作为实体对的特征向量中的一维,由此构成了实体对的特征向量。

c)特征向量降维。利用前面讲到的特征选择算法对上一步构造出的特征向量进行特征选择,将选择出的有效特征重新组成特征向量。

d)构造分类器。用训练语料中实体对降维后的特征向量构造 SVM 分类器。

e)输出分类。利用训练得到的 SVM 分类器判断测试语料中实体对的关系类型。

2　实验结果及其分析

2.1　实验数据

实验使用的数据由 SemEval-2010 评测任务 8 提供。SemEval(Semantic Evaluations)是国际知名的语义处理评测会议,由著名的 ACL(Association for Computational Linguistics)中的 SigLex 组织主办。SemEval-2010 评测任务 8 将实体关系类型分为九类,提供的训练语料包含 8 000 个句子,每个句子均标出了两个实体及其所属关系类型。在本文的实验中,将 8 000 个句子的前 800 个句子作为测试语料,其余的句子作为训练语料。表 2 为实验中训练语料和测试语料的所属关系类型统计。

表 2　语料关系类型统计

关系类型	训练语料	测试语料
other	1 283	122
component-whole	861	82
instrument-agency	448	56
member-collection	617	73
cause-effect	895	108
entity-destination	745	100
content-container	481	59
message-topic	558	76
product-producer	659	61
entity-origin	653	63

2.2　实验评价标准

本文采用准确率 P(precision)、召回率 R(recall)和 $F1$ 值($F1$-measure)作为评测标准。它们的定义如下:

$$P = T/E \quad (4)$$

$$R = T/N \quad (5)$$

$$F1 = 2\times P\times R/(P+R) \quad (6)$$

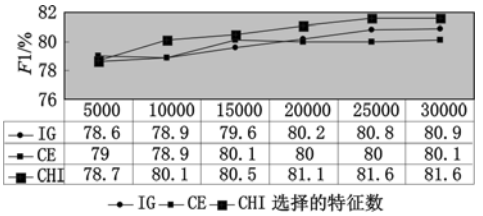
其中:T 为某类被正确分类的实例个数;N 为测试数据中某类实例实际总数;E 为分类器预测为某类的实例总数。

2.3　实验过程及结果分析

实验首先对语料进行词性标注、句法分析和语义角色标注等预处理;然后按照上述特征抽取方法产生特征向量;接着利用上面讲到的特征选择算法进行特征降维;最后,使用 LIBSVM 对抽取出的特征向量进行训练分类。

由于现有特征选择方法通常采用经验方式来确定特征数目,因此为了得到各特征选择方法在达到最佳分类性能时的特征数,本文采用了逐步增加特征数的方法来确定,实验结果如图 1 所示。从图 1 可以看出,对于 IG 方法,特征数从 5 000 增加到 30 000 时,分类性能只增加了 2.3%,即新增的特征并没

有对分类性能产生多大的作用。对于 CE 和 CHI 也是类似的，而且对于 CE 方法，它的分类性能在达到一定程度之后，则不再随着特征数的增加而增加。同时笔者发现，当选择的特征数达到某个阈值时，各特征选择方法性能均会达到最佳状态，如果此时继续增加选择的特征数，性能不但不会进一步提高，而且还有可能下降。对于这个使得性能达到最佳状态的阈值的确定，则需要通过大量实验才能得到。



	5000	10000	15000	20000	25000	30000
IG	78.6	78.9	79.6	80.2	80.8	80.9
CE	79	78.9	80.1	80	80	80.1
CHI	78.7	80.1	80.5	81.1	81.6	81.6

图1　各特征选择方法在不同特征数下的性能比较

表 3 给出了各个特征选择算法对应的实体关系抽取方案的性能比较。

表 3　实体关系抽取方案性能比较

方案	选择的特征		分类性能/%			时间开销/s	
	数量	百分比	*P*	*R*	*F*1	训练	预测
SVM	121 587	100	78.1	85.4	81.6	2 092	935
SVM + IG	30 000	24.7	76.7	85.7	80.9	784	338
SVM + CE	15 000	12.3	75.5	85.4	80.1	565	198
SVM + CHI	25 000	20.1	77.1	86.6	81.6	691	223

比较表 3 的分类性能数据可以发现，无论使用哪一种特征选择方法都没有提高实体关系分类性能，最好的情况也就是不降低它的性能。这是由于在实体关系抽取方案中加入了特征选择算法之后，降低了分类时特征空间维数，而在这个降维过程中，有一些对实体关系抽取有用的信息被丢掉。虽然增加了特征选择的实体关系抽取方案可能会降低实体关系分类性能，但从表 2 的数据可以看出，该类方案依然是有其价值的。这是因为首先这类方案只是略微降低了分类性能，比如 SVM + IG 方案只降低了 0.7%，SVM + CE 方案只降低了 1.5%；其次，该类方案有效地减少了分类时的特征数，提高了效率，比如 SVM + CE 方案以将性能降低 0.7% 为代价将特征数也减少到了 24.7%，而 SVM + CHI 方案则在保持分类性能的基础上将特征数减少到了 24.1%。由此可以看出，该类方案是将分类性能和效率作了一个权衡，在尽量保证分类性能的同时提高分类效率。在实际应用中可以根据需要选择合适的实体关系抽取方案。

对于 IG、CE 和 CHI 三种特征选择方法，从图 1 和表 3 的实验结果可以看出，CHI 是更适合于实体关系抽取的。因为在选择相同特征数时，以 CHI 得到的实体关系抽取性能最好。

3　结束语

由于实体关系抽取问题与文本分类问题的相似性，本文引入了文本分类中的特征选择算法，用于解决基于特征向量的实体关系抽取问题中特征空间维数过高的问题。实验结果表明，本文引入的基于信息增益、期望交叉熵和 χ^2 统计的特征选择算法均能有效地降低实体关系抽取中的特征维数，减少抽取的时间开销，且保持了实体关系抽取的 *F*1 值。然而，特征选择过程希望最好在降低特征维数的同时提高抽取性能，这个目标是困难的，也将是笔者下一步的研究方向。另外，考虑到本文只是简单引入了文本分类中的特征选择算法，下一步也可以组合多个特征选择算法，以期更进一步地进行有效特征降维。

参考文献：

[1] 黄鑫. 基于特征向量的中文实体间语义关系抽取研究[D]. 苏州：苏州大学，2009.

[2] TYMOSHENKO K, GIULIANO C. Semantic relation extraction using Cyc[C]//Proc of the 5th International Workshop on Semantic Evaluation. Stroudsburg, PA: Association for Computational Linguistics, 2010: 214-217.

[3] GIULIANO C, LAVELLI A, PIGHIN D, *et al*. Kernel methods for semantic relation extraction[C]//Proc of the 4th International Workshop on Semantic Evaluations. Stroudsburg, PA: Association for Computational Linguistics, 2007: 141-144.

[4] ZHOU Guo-dong, QIAN Long-hua, FAN Jian-xi. Tree kernel-based semantic relation extraction with rich syntactic and semantic information[J]. Information Sciences, 2010, 180(2010): 1313-1325.

[5] 庄成龙，钱龙华，周国栋. 基于树核函数的实体语义关系抽取方法研究[J]. 中文信息学报，2009，23(1)：1003-0077.

[6] 黄高辉，姚天昉，刘全升. 基于 CRF 算法的汉语比较句识别和关系抽取[J]. 计算机应用研究，2010，27(6)：2062-2064.

[7] LLORENS H, SAQUETE E, NAVARRO-COLORADO B. TimeML events recognition and classification: learning CRF models with semantic roles[C]//Proc of the 23rd International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010: 725-733.

[8] 张彪. 文本分类中特征选择算法的分析与研究题名[D]. 合肥：中国科学技术大学，2010.

[9] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859.

[10] NOVOVIČOVÁ, SOMOL P, HAINDL M, *et al*. Conditional mutual information based feature selection for classification task[C]//Proc of the 12th Iberoamerican Conference on Congress on Pattern Recognition. Berlin: Springer-Verlag, 2007: 417-426.

（上接第 484 页）

[2] 董振东. 知网[EB/OL]. (2010-08-20)[2011-07-06]. http://www.keenage.com/html/c_index.html.

[3] DORIGO M, MANIEZZO V, COLORNI A. Ant system: optimization by a colony of cooperating agents[J]. IEEE Trans on Systems, Man, and Cybernetics-Part B, 1996, 26(1): 1-13.

[4] 段海滨，王道波，朱家强，等. 蚁群算法理论及应用研究的进展[J]. 控制与决策，2004，19(12)：1321-1340.

[5] 吴启迪，汪镭. 智能蚁群算法及其应用[M]. 上海：上海科技出版社，2004：54-58.

[6] 刘群，李素建. 基于《知网》的词汇语义相似度计算[J]. 计算语言学及中文信息处理，2002，7(2)：59-76.

[7] 楼华锋，刘功申. 一种基于段落同现频率的加权方法[J]. 信息安全与通信保密，2009(12)：57-63.

[8] 舒湘沅，杨铭，王延平. 航空项目资源均衡优化问题的蚁群—模拟退火算法[J]. 航空制造技术，2010，18(13)：77-81.

[9] 高尚，汤可宗，杨靖宇. 一种新的基于混合蚁群算法的聚类方法[J]. 微电子学与计算机，2006，23(12)：38-40.

[10] 中国科学院计算技术研究所数字化室. 中文自然语言处理开放平台[DB/OL]. (2011-05-25)[2011-07-06]. http://www.nlp.org.cn/.