中 文 信 息 学 报
第 19 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.19 No.2

编者按 中国中文信息学会信息检索与内容安全专业委员会(筹)于 2004 年 11 月在上海复旦大学成功地召开了"第一届全国信息检索与内容安全学术会议(NCIRCS2004)"。会议的程序委员会从录用的论文中评选出 15 篇优秀论文,本刊编辑部得到授权,在此发表,以飨读者。

文章编号:1003 - 0077(2005)02 - 0001 - 06

实体关系自动抽取

车万翔,刘 挺,李 生

(哈尔滨工业大学 计算机学院,黑龙江 哈尔滨 150001)

摘要:实体关系抽取是信息抽取领域中的重要研究课题。本文使用两种基于特征向量的机器学习算法,Winnow 和支持向量机(SVM),在 2004 年 ACE(Automatic Content Extraction)评测的训练数据上进行实体关系抽取实验。两种算法都进行适当的特征选择,当选择每个实体的左右两个词为特征时,达到最好的抽取效果,Winnow 和 SVM 算法的加权平均 F-Score 分别为 73.08%和 73.27%。可见在使用相同的特征集,不同的学习算法进行实体关系的识别时,最终性能差别不大。因此使用自动的方法进行实体关系抽取时,应当集中精力寻找好的特征。

关键词:计算机应用;中文信息处理;实体关系抽取;ACE 评测;特征选择

中图分类号:TP391 **文献标识码**:A

Automatic Entity Relation Extraction

CHE Wan-xiang, LIU Ting, LI Sheng

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Entity Relation Extraction is an important research field in Information Extraction. Two kinds of machine learning algorithms, Winnow and Support Vector Machine (SVM), were used to extract entity relation from the training data of ACE (Automatic Content Extraction) Evaluation 2004 automatically. Both of the algorithms need appropriate feature selection. When two words around an entity were selected, the performance of the both algorithms got the peak. The average weighted F-Score of Winnow and SVM algorithms were 73.08% and 73.27% respectively. We can conclude that when the same feature set is used, the performance of different machine learning algorithms get little difference. So we should pay more attention to find better features when we use the automatic learning methods to extract the entity relation.

Key words: computer application; Chinese information processing; entity relation extraction; ACE evaluation; feature selection

1 引言

随着计算机的普及以及互联网的迅猛发展,大量的信息以电子文本的形式出现在人们面前。为了应对信息爆炸带来的挑战,迫切需要一些自动化的工具帮助人们在海量信息源中迅

收稿日期:2004 - 06 - 20

基金项目:国家自然科学基金资助(60435020)

作者简介:车万翔(1980—),男,黑龙江人,博士研究生,主要研究领域为自然语言处理,信息检索.

速找到真正需要的信息。信息抽取(Information Extraction)研究正是在这种背景下产生的。信息抽取的主要目的是将无结构的文本转化为结构化或半结构化的信息,并以数据库的形式存储,供用户查询以及进一步分析利用。信息抽取系统的主要功能是从文本中抽取出特定的事实信息,我们称之为实体(Entity)。例如:时间(TIME)、组织机构(ORG)、人物(PER)以及武器(WEAPON)等等。

然而,在大多数的应用中,不但要识别文本中的实体,还要确定这些实体之间的关系,我们称其为实体关系抽取。与实体抽取类似,实体关系的类型也是预先定义的,例如:地理位置关系(PHYS)、雇佣关系(EMP-ORG)等等。假设文本中提到"…美国第七舰队司令…",其中"美国第七舰队司令"和"美国第七舰队"分别为人物(PER)和组织(ORG)实体。而它们又构成了一种雇佣关系(EMP-ORG),即"美国第七舰队司令"受雇于"美国第七舰队"。通过以上介绍,我们发现如果说信息抽取的主要功能是自动将文本转化为数据表格,实体抽取确定了表格中各个元素的话,实体关系抽取则是确定这些元素在表格中的相对位置。可见,实体关系抽取是信息抽取中的重要环节。

近年来,随着实体抽取研究逐步实用化,实体关系抽取的研究也越来越受到人们的重视。其中美国国防高级研究计划委员会(DARPA, the Defense Advanced Research Projects Agency)资助的 MUC(Message Understanding Conference)会议于 1998 年最后一次 MUC - 7[1] 上首次引入了关系抽取(模板关系,Template Relation)任务。

随着 MUC 会议的停办,美国国家标准技术研究院(NIST)组织了自动内容抽取(ACE, Automatic Content Extraction)评测*,它从 1999 年开始继续进行信息抽取方面的评测。ACE 评测 1999 年 7 月开始酝酿,2000 年 12 月正式开始启动,迄今已经举办过四次评测(2000 年 5 月、2002 年 2 月、2002 年 9 月、2003 年 10 月),最近正在进行第五次评测(2004 年 8 月)。其研究的主要内容是自动抽取新闻语料中出现的实体、关系、事件等内容。目前 ACE 评测主要有两大任务:实体识别(EDR, Entity Detection and Recognition)和关系识别(RDR, Relation Detection and Recognition)。其中,RDR 定义了较为详细的关系,包括 7 个大类和若干子关系**。ACE 评测提供的语料不仅是英文,还包括中文和阿拉伯文。在此,我们只研究中文实体关系抽取问题。

通常,人们将关系抽取问题转化为一个分类问题,即首先列出一个句子中所有的实体对,然后使用一个分类器决定哪些是我们真正需要的关系。

和分类问题通常的解决办法一样,人们最初也是使用知识库[2]的方法来解决该问题。但方法需要专家构筑大规模的知识库,这不但需要有专业技能的专家,也需要付出大量劳动。

为了克服知识库方法的缺点,人们后来使用机器学习的方法来解决此问题[3,4]。该方法不需要有专业技能的专家书写知识库,只需要有一定专业知识的人对任意两个实体之间的关系做出是与不是我们需要的关系的判断即可。然后以此为训练数据,使用各种学习方法构造分类器。

通常的机器学习算法需要构造特征向量形式的训练数据。然后使用各种机器学习算法,如支持向量机(SVM)[5]、Winnow[6]等作为学习机构造分类器。这种方法被称作基于特征向量的学习算法。

接着又出现了基于 Kernel 的学习算法,它最早在支持向量机(SVM)方法中被引入,后来发

---

* http://www.nist.gov/speech/tests/ace/index.htm

** http://www.ldc.upenn.edu/Projects/ACE/

现多种学习方法可以使用 Kernel 的形式来表示。它们又被称作基于 Kernel 的学习算法。在自然语言处理领域应用基于 Kernel 的学习算法[7,8]，与基于特征向量的学习算法不同，其不需要构造特征向量，而是直接使用字符串的原始形式作为处理对象，需要做的只是计算任何两个对象之间的 Kernel (Similarity)函数。Zelenko 等[9]以及 Culotta 等[10]使用 Kernel 的方法解决关系抽取问题，取得了较好的结果，他们使用的方法需要对处理对象进行浅层的句法分析。然而，Kernel 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学习算法作为实体关系抽取的方法。

文章的第 2 部分简单介绍了基于特征向量的算法的基本概念和学习过程。在第 3 部分中，介绍了在实体关系抽取中所采用的多种特征向量的构造方法。第 4 部分给出了试验的结果以及对于结果的一些讨论。最后是本文的结论和对未来工作的展望。

2 基于特征向量的机器学习算法

下面，我们就基于特征向量的机器学习算法加以简单介绍。所谓特征向量，是实例的一种数值化的表示方式。也就是说，一个实例被转化为特征向量 x，其中 x_i 为 N 维特征向量 x 的第 i 个元素。基于特征向量的机器学习算法的目的就是对于给定的一组训练数据 (x^1, y^1)，$(x^2, y^2), \cdots, (x^n, y^n)$，其中对于二元分类问题 $y^i \in \{-1, +1\}$，学习一个分类函数 f，使得对于给定新的特征向量 x'，f 能够将其正确的分类，即 $f(x') = y'$。

分类函数 f 一般定义为一个由权值向量 ω 决定的超平面，其能够将标号为 -1 和标号为 +1 的数据严格分开。权值向量 ω 的求解方法由各种机器学习算法完成。常用的有在线学习算法(Online Learning)和支持向量机(SVM)等。

在线学习算法的基本思想是对于权值向量 ω 和一个新的训练数据 x^i，如果权值向量 ω 对应的超平面不能将 x^i 正确地分开，就可利用 x^i 来修正 ω。可对训练数据反复迭代这一过程，直至所有的训练数据都能正确分开(如果是线性可分)。Winnow[6] 和 Perceptron[11] 等算法是常用的在线学习算法。

支持向量机最早由 Vipnik 提出，后成功的应用于文本分类[12]等自然语言处理领域。支持向量机的基本思想是使 ω 构成的超平面分割训练数据能够获得最大的边缘(Margin)。

对于自然语言处理问题，如何构造特征向量成为使用基于特征向量学习算法的一个重要环节。例如，在文本分类任务中，通常使用一个词表作为特征向量，而向量中元素的值可以是二元的 1 或 0，代表某个词出现与否，或者是该词在一篇文档中出现的次数，目前使用词的 tf × idf 值作为元素值取得了较好的分类效果[13]。在其它一些自然语言处理问题中，向量的每个元素则是一些预先定义的特征在实例中出现与否。即根据特征函数 $f_i: H \times T \rightarrow \{0,1\}$ 决定第 i 维向量元素的值。其中 H 是实例上下文的集合，T 是实例所属类别的集合，则特征向量的第 i 维向量元素 $x_i = f_i(h, t)$。例如在词性标注问题中，我们可能有下面的一个特征：

$$x_{100} = f_{100}(h, t) = \begin{cases} 1 & \text{if 当前词 } w_i \text{ 是“打”并且 } t = V \\ 0 & \text{其它} \end{cases}$$

于是，便构成了一个维数巨大的特征向量。此特征向量即可作为某一机器学习算法的输入数据，进行学习和预测。

上面我们简单的介绍了几种机器学习算法及一般特征向量的构造过程。接下来的部分我们将详细介绍在实体关系抽取中，我们所使用的特征向量构造方法。

3 实体关系抽取中的特征向量构造

通过上面部分的介绍，我们知道为了使用基于特征向量的机器学习算法进行实体关系抽取，最重要的过程是实例特征向量构造。选取恰当的特征能够较好的对实体进行表述，有利于学习效果的提高。所谓恰当的特征即是与分类相关的特征，具有较强的区分度（distinguishing）。仍然以词性标注问题为例，通常人们选择一个词周围 w（w 一般不是很大）个词和词性及其组合作为特征，因为一般情况下，距离较远的词对于词性标注不起重要作用，而且如果 w 的选择过大，会增加大量的计算。

目前，我们只考虑一个句子中的两个实体之间的关系，而不考虑跨越句子的实体之间的关系。也就是说实体关系抽取问题的输入文本是一个句子以及句子中被正确识别的实体，所谓正确识别，不但要识别出实体的类别，还要正确确定实体的边界。根据 ACE 对实体的定义，共分七大类，每个类别又有若干个子类。例如对于输入的原始句子“南国国会大厦是抗议的主要中心。”，其中已知含有“南国国会（TYPE = “ORG”，SUBTYPE = “Government”）”和“南国国会大厦 TYPE = “FAC”，SUBTYPE = “Building”“两个实体。经过实体关系的抽取，确定“南国国会”和“南国国会大厦”的关系类型为 TYPE = “ART”，SUBTYPE = “User-or-Owner”。

通过对训练数据的分析，我们发现，实体的类型对确定它们之间的关系有很好的作用。另外两个实体在句子中出现的顺序（或者包含关系）也是决定它们之间关系的重要因素。实体周围的 w 个词也是较好的特征。例如对于实体“国会外的示威人群（TYPE = “PER”）”和“国会（TYPE = “FAC”，SUBTYPE = “Building”）”所形成的 TYPE = “PHYS”，SUBTYPE = “Located”关系，“国会”右面的“外”有着很强的指示作用。

综上，我们为一个句子中任意实体对（$E1$，$E2$）构造的特征向量如图 1 所示：

E1.TYPE，E2.TYPE，E1.SUBTYPE，E2.SUBTYPE，Order，

Wi－w，Wi－w－1，…，Wi－1，Wi＋1，…，Wi＋w－1，Wi＋w，ti－w，ti－w－1，…，ti－1，ti＋1，…，ti＋w－1，ti＋w，

Wj－w，Wj－w－1，…，Wj－1，Wj＋1，…，Wj＋w－1，Wj＋w，tj－w，tj－w－1，…，tj－1，tj＋1，…，tj＋w－1，tj＋w。

图 1 特征向量的构造

其中，E.TYPE 为实体所属的大类，E.SUBTYPE 为实体所属的子类。Order 为两个实体之间的位置关系，分别为 0（表示 $E1$ 位于 $E2$ 的左面），1（表示 $E1$ 位于 $E2$ 的右面），2（表示 $E1$ 包含 $E2$）和 3（表示 $E2$ 包含 $E1$）。而 i 和 j 分别为先后出现的两个实体的位置。W_k 和 t_k 分别为位置 k 处的汉语词和词性。

将句子中任意两个实体对的这些所有的属性构成的向量看成 x^i，其分类标记看成 y^i，则构成多元分类的一个样例（x^i，y^i）。多元分类可用二元分类器来完成，有“一对多”和“两两”分类方法，我们用“一对多”方法。

4 试验结果以及讨论

4.1 实验数据及评测指标

我们使用 2004 年 ACE 评测的训练数据作为实验数据。数据的来源为广播新闻（Broadcast News，314 篇），中文树库（Chinese Tree Bank，106 篇）和新华社新闻（XinHua News，226 篇）。并随机选取其中的 1/3 作为测试数据，其余的 2/3 作为训练数据。ACE 评测的训练数据，不但标注实体以及实体的各种属性，还标注了实体关系以及关系的属性。数据以及标注结果以 XML 格

式存储。

每个句子中的任意两个实体即形成一个实例，表 1 列出了所有实例的统计信息，其中我们只考虑七个大类的分类情况。

表 1　实例统计信息

类　别	ART	DISC	EMP-ORG	GPE-AFF	OTHER-AFF	PER-SOC	PHYS	NO-Relation	总　数
数　量	160	589	2673	1368	47	217	1117	15048	21219

对于实体关系抽取问题的性能评测，我们可以使用信息检索（Information Retrieval）中的评测方法，采用 F－Score 对最终系统的性能进行评价。其定义如下：

$$\text{F} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

其中准确率（Precision）和召回率（Recall）的定义为：

$$\text{Precision} = \frac{\text{某类被正确分类的实例个数}}{\text{分类器预测的某类实例总数}}$$

$$\text{Recall} = \frac{\text{某类被正确分类的实例个数}}{\text{测试数据中某类实例总数}}$$

对于多类实体的抽取问题，我们最终使用各类性能的加权平均值作为最终的评测，以计算加权平均 F－Score 为例，其公式如下：

$$F_{\text{avg}} = \frac{\sum_i C_i \times F_i}{\sum_i C_i}$$

P_{avg}与 R_{avg}的定义类似。其中 C_i 为第 i 类的实例的个数，F_i 为第 i 类的 F-Score 值。

4.2　实验结果

在构造实例向量的时候，我们首先对原始数据进行分词和词性标注处理。分词的主要难点是名、地名等未登录词的处理，而我们根据 ACE 训练语料中对实体已知的标注信息，可以很好的回避这一难题，大幅度提高分词的准确率。同时词性标注的性能也有所提高。

如第 2 部分所述，在构造特征向量的时候，随着窗口大小 w 的不同，我们得到不同的特征向量构造方法。使用 Winnow 算法和 SVM 两种机器学习算法进行训练和识别，得到表 2 所示的结果。其中，Winnow 算法使用 UIUC 大学的 SNow 算法库*，迭代次数设置为 3000 次。SVM 算法使用台湾大学的 libsvm 算法库**，并使用默认的参数。

表 2　实体关系识别结果

算法及特征抽取方法		P_{avg}	R_{avg}	F_{avg}
Winnow	w = 0	59.93%	60.65%	57.05%
	w = 1	72.88%	68.54%	70.39%
	w = 2	74.75%	71.69%	73.08%
	w = 3	73.54%	71.30%	72.29%
SVM	w = 0	66.20%	59.68%	60.60%
	w = 1	72.99%	69.29%	70.89%
	w = 2	76.13%	70.18%	73.27%
	w = 3	74.22%	69.31%	71.84%

通过对表 2 的分析，我们发现在 $w = 2$ 时，两种算法同时获得最优的结果。而且 SVM 算法较之 Winnow 算法的结果要好。但是，在训练和测试的时候，SVM 算法都需要较长的运行时间。

---

* http://l2r.cs.uiuc.edu/cogcomp/software/snowreg.html

** http://www.csie.ntu.edu.tw/cjlin/libsvm/

5 结论

本文介绍了信息抽取研究领域中逐渐为人们所重视的实体关系抽取研究工作。并以2004年ACE评测的训练数据作为实验数据，使用两种基于特征向量的机器学习算法，Winnow和SVM进行实体关系抽取实验。两种算法都进行适当的特征选择，当选择每个实体的左右两个词为特征时，达到最好的抽取效果。可见，实体的左右两个词是进行实体关系抽取的较好的特征，但是若选择的词过多，会引入不相关特征，降低系统的性能。

而在使用相同的特征向量，不同的学习算法进行实体关系的识别时，最终性能差别不大，因此我们需要根据自己的需求选择算法。即当对学习的效率要求不高而对最终的学习性能要求较高时，可以选择SVM算法；相反，若对学习的效率要求很高而对最终的学习性能要求不高时，可以选择Winnow算法。另外，特征的选择对最终系统的性能影响较大。因此在使用机器学习的算法进行分类等应用时，应尽量根据领域知识，选择相关特征。

因此我们下一步的工作就是寻找更好的特征，例如加入句法、词的语义等信息，构造特征丰富的特征向量，进一步改善实体关系抽取的性能。

参 考 文 献：

[1] In: Proceedings of the 6th Message Understanding Conference (MUC-7)[C]. National Institute of Standars and Technology, 1998.

[2] C. Aone and M. Ramos-Santacruz. Rees: A large-scale relation and event extraction system[A]. In: Proceedings of the 6th Applied Natural Language Processing Conference[C], pages 76-83, 2000.

[3] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel, and the Annotation Group. Algorithms that learn to extract information-BBN: Description of the SIFT system as used for MUC[A]. In: Proceedings of the Seventh Message Understanding Conference (MUC-7)[C], 1998.

[4] S. Soderland. Learning information extraction rules for semi-structured and free text[J]. Machine Learning, 1999. 34 (1-3):233-272.

[5] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines[M]. Cambridge University Press, Cambirdge University, 2000.

[6] T. Zhang. Regularized winnow methods[A]. In: Advances in Neural Information Processing Systems 13[C], pages 703-709, 2001.

[7] D. Haussler. Convolution kernels on discrete structures[R]. Technical Report UCSC-CRL-99-10, 7, 1999.

[8] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels [R]. J. Mach. Learn. Res., 2:419-444, 2002.

[9] D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relation extraction[R]. J. Mach. Learn. Res., 3: 1083-1106, 2003.

[10] A. Culotta and J. Sorensen. Dependency tree kernels for relation extraction[A]. In: Proceedings of ACL[C]. 2004. Barcelona, Spain.

[11] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm[J]. In: Computational Learning Theory, pages 209-217, 1998.

[12] T. Joachims. Text categorization with support vector machines: learning with many relevant features[A]. In: C. Nédellec and C. Rouveirol, editors, Proceedings of ECML-98, 10th European Conference on Machine Learning [C], number 1398, pages 137-142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[13] K. Aas and L. Eikvil, Text categorization: A survey, tech. rep.[R], Norwegian Computing Center, June 1999.