交通工具名识别系统的设计与实现*

王振宇[13] 谭红叶[123] 郑家恒[13] 张虎[13]

([1]山西大学计算机与信息技术学院，山西 太原 030006)

([2]哈尔滨工业大学计算机科学与技术学院 黑龙江 哈尔滨 150001)

([3]计算智能与中文信息处理教育部重点实验室 山西 太原 030006)

Email:zhenyu_wangsxdx@126.com

摘要： 交通工具名的正确识别对信息抽取、自动问答等信息处理任务意义重大，为了解决获取标注语料困难这个问题，本文实现了一种基于Bootstrapping的交通工具名识别方法，其特点是：（1）通过手工标记小部分语料逐渐学习得到大量标注信息，（2）其中评价模式和样例时采用了计算信息熵增益的方法，更加精确的得到它们的度量方法。在ACE语料上进行测试，实验表明该方法在交通工具名识别中是有效的。

关键词： 交通工具名识别，Bootstrapping，信息熵增益，相似度计算

the Identification of Vehicle Based on Bootstrapping

WANG Zhen-yu[ac] TAN Hong-ye[abc] ZHENG Jia-heng[ac] ZHANG Hu[ac]

([a] Department of Computer and Information Technology, Shanxi University, Taiyuan 030006)

([b]Department of Computer Science and Technology, Harbin Institute of Technology，Harbin 150001)

([c]Key laboratory of education for computation intelligence and Chinese information processing, Taiyuan ,030006 China)

Email:zhenyu_wangsxdx@126.com

Abstract: The correct identification of vehicle play a important role in information extraction, automatic quizzes, and other information processing tasks, in order to solve this difficulty of obtaining the annotated corpus, the paper developed a method of identification of vehicle based on Bootstrapping, the main ideas are (1)manually marking small part of the corpus, then Learn substantial tagging information gradually, (2)the evaluation of sample and pattern uses the calculation of information entropy plus, a more precise measurement method for them. We test our method on ACE corpus, the results show that the method of vehicle identification is effective.

Keywords: identification of VEH, Bootstrapping, information entropy plus, similarity calculation

1 引言

随着互联网的不断发展，信息呈爆炸式增长，如何从非结构化的普通文本中自动抽取信息构建知识库是必须解决的问题。命名实体识别（Named Entity Recognition）是信息抽取的主要任务，实体识别是五个ACE识别任务中的重要一个，其中，交通工具名属于ACE实体类别的一类，因此交通工具名识别是ACE项目中重要的组成部分。

交通工具名的正确识别对信息抽取、自动问答等信息处理任务意义重大。例如，在某些新闻事件中，交通工具名是其重要特征；在一些问答系统如旅游问答系统中，交通工具名对人们有一定的指导作用。

目前专门针对交通工具名识别的方法还不多，已有方法大部分是对语料中多种扩展类别的

*基金项目：国家自然科学基金（60473139，60775041）项目。

作者简介：王振宇（1984-），女，山西平遥人，硕士研究生，研究方向为自然语言处理。

名实体进行识别，刘非凡，赵军等人采用有监督的学习方法进行产品名的识别[1]，但是有监督学习需要足够的训练语料，为了克服获取熟语料的困难，Roman Yangarber,Winston Lin,Michael Thelen等人采用Bootstrpping算法进行基于模式的实体类别研究[2,3,4]，但是这些方法使用的模式单一，在模式匹配的时候要求精确匹配，所以取得的性能比较有限。

本文致力于研究采用Bootstrapping方法进行交通工具名识别，通过相似度计算来识别交通工具名，避免了模式的精确匹配，而且可以根据交通工具名自身的特点添加一些规则，大大提高了性能，在ACE语料上进行测试取得了不错的效果。

2 相关概念

2.1 相关定义

1）上下文模式

上下文模式是指文本中表达关系和事件信息的重复出现的特定语言表达形式,可以按照特定的规则通过模式匹配，触发抽取特定信息。上下文模式被看作是由项组成的有序序列，每个项对应于一个词（或者词组）的集合。

设上下文模式为P，它可以表示为：

$$P=(pattern1,pattern2,\cdots,patterni,\cdots)$$

其中，$patterni$表示根据特征模板抽取的第i个模式。

2）模式匹配

模式匹配(pattern-matching)：系统将输入的句子同模式进行匹配，根据匹配成功的模式，得到相应的解释。

如：对于当前模式“TLUnigram 乘坐”，分别考虑语料中每一个句子，若有一句话为：

乘坐/ 大客车/ 的/ 警察/ 强行/ 进入/ 这个/ 地方/ 以后/ ，/

则匹配成功，且抽取出词语“大客车”。

3） 种子，样例，实例

种子是在Bootstrapping方法中借助的少量“指导”，人工标注少量的交通工具名作为种子。

样例是Bootstrapping方法中，经过抽取种子的上下文模式，然后再与原文模式匹配后，抽取出的词语。

实例是在测试语料中，要进行类别判断的词语。

4） 特征向量

特征向量是表示一类实体或者一个实例所有特征的向量，在交通工具名识别系统中涉及到两种特征向量，分别为类特征向量和实例特征向量。

5） 归一化频数

归一化频数是用来表示特征向量中各项的权重的。

类特征项的归一化频数等于该特征项在训练语料中的出现次数与该类特征向量中特征项的总个数的比值。

实例特征项的归一化频数等于该特征项在待测实例所在当前句中的出现次数与当前句中特征项总个数的比值。

3 交通工具名识别系统的设计

3.1系统设计的主要思想

基于bootstrapping的交通工具名识别算法的主要思想是，将交通工具名的特征信息和待识别实例分别表示为类特征向量和实例特征向量，通过计算它们的相似度来判断待识别实例是否属于交通工具名，在类特征向量的获取过程中采用Bootstrapping算法，其中模式和样例评价时采用计算信息熵增益的方法。

3.2系统结构

基于Bootstrapping方法的交通工具名识别算法涉及到三个数据集：验证集、训练集和测试集。各数据集与各算法的对应关系如图1所示：

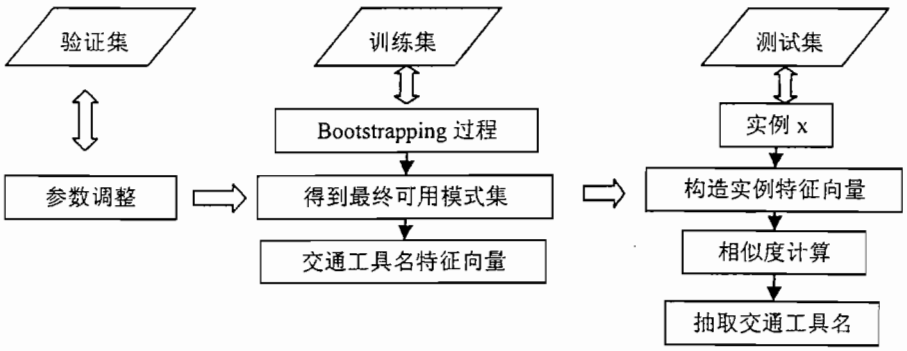

图1 基于Bootstrapping方法的交通工具名识别系统结构图

3.3信息熵增益

设$T=(U,C\cup D,V,f)$是一个决策表，且$R\subset C$。则对于任意属性$a\in C-R$的信息熵增益$SGF(a,R,D)$定义为[5]：

$$SGF(a,R,D)=H(D|R)-H(D|R\cup\{a\})$$

其中：论域U是所有的种子和样例，D是决策属性，即属于交通工具名与否。C是条件属性，即所有的模式。R是核，即已标注的交通工具名的部分模式。a是C中除R以外的模式。$H(D|R)$是D相对于R的条件熵，对所有R分类下不确定的等价类进行再划分的熵。它体现了用R对论域U划分所形成的结果的不确定性。

$SGF(a,R,D)$的值越大，说明在R已知的条件下，属性a对于决策D就越重要，因此把$SGF(a,R,D)$作为模式和样例评价时的标准。

4、系统实现

4.1 Bootstrapping算法

Bootstrapping算法是典型的一种弱指导学习算法，主要思想就是利用一定的种子，经过反复迭代得到一定规模的弱标注语料和上下文模式集合。

针对此算法的特点，选择可变数组ArryList作为存储结构，在此算法的实现过程中，定义了如下变量：

ArrayList pattern_use　　最终可用模式集；
ArrayList pattern_all　　可用模式集；
ArrayList entities　　识别出的样例（最初有重复的）；
ArrayList ne　　识别出的实体（去重后）；
ArrayList instance　　识别出的实例（要加入种子集的）；

该算法流程图如图2所示。

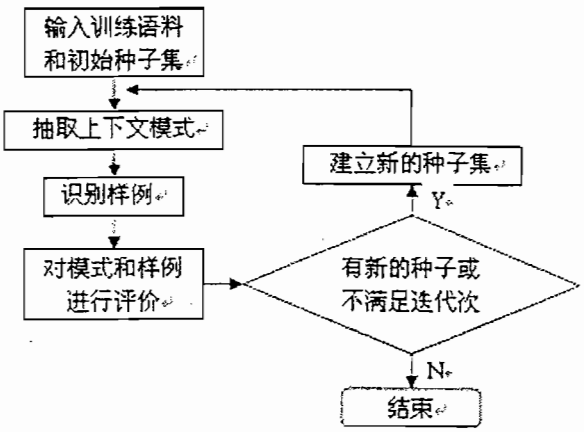

图2 Bootstrapping算法流程图

在Bootstrapping算法中模式的评价非常重要。因为利用不可靠的模式识别出的错误实例会被作为新的种子再次进行模式的抽取，这样所抽取的模式必定是错误模式。而模式错误又会被不断传递放大。因此需要在每一次迭代过程中对模式和识别出的实例进行评价。只有满足一定分值的实体名和模式才能进行循环。本文采用计算信息熵的增益来实现新样例和模式的评价。

4.2 特征向量的构造

类特征向量的构造：Bootstrapping过程结束后，得到一个最终可用模式集合，该集合中每

一个元素代表了交通工具名的特征，分别计算它们的归一化频数，并排序，选择出现两次以上的特征项作为类特征向量中的特征项，对应的归一化频数作为该特征项的频数。

实例特征向量的构造：对于一个待识别实例，抽取其上下文模式，每一个模式代表了其特征，分别计算它们的归一化频数，并排序，将这些特征项作为实例特征向量中的特征项，对应的归一化频数作为该特征项的频数。

4.3 交通工具名识别

交通工具名识别是通过计算类特征向量和实例特征向量的相似度来进行的。由于类特征向量维数很大，而实例特征向量维数与其相差很大，类特征向量的降维不易实现，所以这里不采用以往的计算向量夹角余弦值的方法来得到。本文所用的方法是直接计算两个向量的“类内积”，即先比较特征，若特征相同，则将频次相乘，最后将所有的积相加，用公式表示为：

$$sim(\vec{x},\vec{y})=\sum_{j=1}^{n}\sum_{i=1}^{m}af_{y_j}f_{x_i}$$

其中，

$\vec{x}=(x_1,x_2,\cdots,x_m),\vec{f}_x=(f_{x_1},f_{x_2},\cdots,f_{x_m})$，在本文中，$\vec{x}$ 表示类特征向量，x_i 表示特征项，f_{x_i} 表示特征项对应的归一化频数，$\vec{y}=(y_1,y_2,\cdots,y_n),\vec{f}_y=(f_{y_1},f_{y_2},\cdots,f_{y_m})$，$\vec{y}$ 表示实例特征向量，y_i 表示特征项，f_{y_i} 表示特征项对应的归一化频数。

$$a=\begin{cases}0,若x_i\neq y_i\\1,若x_i=y_i\end{cases}$$

计算出相似度之后，根据阈值大小判断此实例是否属于交通工具名。

5、实验结果及分析

5.1、实验建立

数据集。我们采用ACE语料来进行实验。用哈工大的分词工具进行分词，将其分为三份：训练集、验证集、测试集。表1 给出了这三个集合的具体情况：

数据集	字数	分句数	交通工具名个数
训练集	18.9万	9155	370
验证集	6万	2960	128
测试集	7.2万	3827	136

参数的调整。我们利用验证集进行参数调整，在整个实验过程中，涉及到四个参数，分别是：bootstrapping迭代次数、每次迭代加入最终可用模式集的候选模式个数、每次迭代后选择样例

的个数和交通工具名相似度阈值。经过多次实验，我们最终选择这四个参数分别为：5、20、10

测试过程：对于一个待识别词，首先构造它的实例特征向量，同样抽取窗口长度为3的上下文模式，每个特征在本句的频率作为其权重。计算两个向量的相似度，若相似度大于阈值，则判断其为交通工具名。在我们的实验中，相似度阈值为0.001。

冲突消解。由于交通工具名可能包含不止一个分词单位，如在句子

愤怒的塞族人还发火焚烧了</一辆联合国警车/>VEH，

这句话中交通工具名包含4个分词单位，所以应该事先统计训练语料中的交通工具名的最大长度，在测试时就应该判断组合词是否属于交通工具名。这样就会产生冲突，当遇到不同长度的词均满足阈值条件时，选择相似度最大的确定为交通工具名。

5.2、实验结果

对命名实体识别的评价采用三个指标：准确率P、召回率R和F值

实验结果如表2：

P	R	F
36.8%	89.5%	52.1%

实验结果表明，在样例评价时采用计算信息熵增益的方法效果比较理想，因为信息熵增益可以反映一个新的样例对整个交通工具名集合的影响程度。

在实验过程中，我们发现有很多错误识别的例子，如：艺术节、总统府等，若将它们视为交通工具名的停用词，实验效果有很大改善，准确率为41.1%，召回率为90.2%，F值为56.5%。

本文实现了一种基于Bootstrapping的交通工具名识别系统，通过Bootstrapping算法将交通工具名的有效特征表示为类特征向量，将待测实例的特征表示为实例特征向量，通过计算两个向量的相似度判断一个词是否属于交通工具名。模式和样例评价是Bootstrapping算法中的重要一步，我们通过计算信息熵增益来度量他们的重要程度。这种方法克服了依赖大量标注语料的缺点，在ACE语料上进行实验，取得了满意的结果。目前的实验还有以下方面需要改进：（1）种子的选取，目前只是随随机人工选取，有些不具有代表性，大大影响了识别效果；（2）参数的调整。

参考文献

[1] 刘非凡,赵军等.面向商务信息抽取的产品命名实体识别研究.第八届全国计算语言学联合学术会议(JSCL-05), 南京, 2005.

[2] Roman Yangarber, Winston Lin, Ralph Grishman. Unsupervised Learning of Generalized Names.In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002).

[3] Winston Lin, Roman Yangarber, Ralph Grishman. Bootstrapped Learning of Semantic Classes from Positive and Negative Examples. *Proc. ICML2003 Wkshpth on The Continuum from Labeled to Unlabeled Data*, 2003.

[4] Michael Thelen, Ellen Riloff. A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, New Brunswick, NJ: Association for Computational Linguistics. 2002

[5] 王国胤，于洪，杨大春.基于条件信息熵的决策表约简.计算机学报，2002，7（25）:759-766