基于组合特征的关系抽取研究

徐 谦[1] 祝晓明[2]

(1.南京大学计算机科学与技术系,江苏 南京 210093;2.南京政治学院,江苏 南京 210003)

摘 要:实体关系抽取是信息抽取领域的一个非常重要的子领域。实体关系抽取任务主要是利用计算机自动抽取命名实体之间的关系。现有的关系抽取方法有基于规则的方法,基于特征向量的方法和基于核的方法。本文主要介绍了一种利用了多种特征进行组合来进行关系抽取的方法,并结合ACE语料进行了实验验证了组合特征的有效性。

关键词:关系抽取;特征向量;svm 分类器

中图分类号:TP311.132 文献标识码:A 文章编号:1672-3198(2009)02-0277-02

1 引言

当今社会中充斥着大量的信息,一个最大的信息源就是万维网(World Wide Web)。例如:Google.com 到现在为止已经索引了超过 45 亿的页面,包括 html,PowerPoint,PDF 和其它格式的文件,这些文件中有着巨量的自然语言文本。这类文本的数量非常巨大。如何让计算机帮助人们获取这些巨量文本中的信息?计算机如何能提取出其中有用的信息?信息抽取(Information Extraction)研究正是在这样的背景下产生的。信息抽取的主要目标是将无结构化的自然语言转化为结构化的或者半结构化的信息,并以数据库或者其他易于查找的方式存储起来。

信息抽取的主要功能是从文本中抽取特定的信息,称之为实体(Entity)。同时,在很多应用中,我们不但要识别语言中不同的实体,还要识别实体之间不同的关系,这就是实体关系抽取。

2 现有的关系抽取方法

所谓实体关系抽取,就是根据给定的两个实体和对应的上下文,自动地判断出这两个实体之间的关系。其中,实体(entity)是客观世界中的一个存在物,如:一个人,一个组织,一件物品等等。在我们实体关系抽取问题中还经常提到另一个概念就是实体的具体存在(mention),这个是用来区别同一个实体的不同出现,如:李明今天不在家,他去外面踢球了。其中,李明 和 他 都是指的同一个实体 李明,但它们是这个实体的两个不同的具体存在(mention)。而实体关系抽取的任务是要抽取不同实体之间的关系,例如:郭士纳是 IBM 公司的主席。(Gerstner is the chairman of IBM Corporation.),这句话中,我们已知的两个实体是郭士纳(PER)和 IBM 公司(ORG),他们之间的关系是一种雇用关系(person-affiliation)。现有的关系抽取方法主要有 3 种:基于规则的方法,基于特征向量的方法,基于核的方法。

基于规则的关系抽取方法往往通过引入大量的语言学知识来描述各种关系模式。如 Miller 等采用集成的语法分析方法从文本中抽取关系,该方法通过使用一个生成性的概率模型将所有的实体、关系和语法等决策等同时集成到一棵语法分析树中。这类方法的主要问题是语言规则的建立比较困难,需要专业人员和很长的时间。

基于特征向量的关系抽取方法是通过大量的各种语言特征,例如词法、语法、语义等特征来表示关系实例。Zhou Guodong 等基于 SVM 模型利用词汇的、语法的和语义的等各种知识实现了基于特征的关系抽取系统,取得了较好的实验效果。基于特征向量的关系抽取方法的主要问题是特征选取过程比较零散,不能保证抽取的数字特征包含了所有的信息,缺乏一种系统化的完备的抽取特征的方法。

基于核的关系抽取方法通过将文本数据映射到高维空间来增强线性分类器的计算能力。基于核方法的关系抽取研究主要着重于设计各种核函数来利用语言中的相关的特征。Zelenko 等最早基于浅层句法分析树设计了一个树核来表示两个关系实例之间的相似性;Culotta 等则扩展了 Zelenko 的思想,在增量的依存分析树上计算两个关系实例的相似度。基于核的方法的主要问题在于核的计算过程时间很长。

3 特征向量的构造

实体关系抽取的主要任务是:给定一个自然语言片段,其中包含了两个标记的实体,根据上下文信息,在一组已经预先定义好了的语义关系类型中,找到这两个实体属于哪一种关系。形式化的表示如下:$r=(s, arg_1, arg_2)$ 表示了一个实体关系实例,其中,s 是一个语言片段(一般是一个句子),arg_1 和 arg_2 是两个实体,规定 的位置在 s 中先于 arg_2。给定一组关系实例$\{r_i\}$,每个实例都有个标记 r_i , 是一组预先定义好的实体关系类型的集合,其中包含了一种类型 nil(表示两个实体之间不存在关系),我们的目标就是要学习一个函数 $f:\{r_i\}$,把任意一个实体关系实例 r 映射到一个类型 t 。

所谓特征向量,是实例的一种数值化的表示方式。也

作者简介:徐谦,南京大学计算机科学与技术系 2006 级硕士研究生;祝晓明,男,南京政治学院基础系,副教授,研究方向:计算机应用领域的教学研究。

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) , , (x^n, y^n) , 其中对于二元分类问题 y^i {- 1, + 1}, 学习一个分类函数 f, 使得对于新的特征向量 x , f 能够将其正确的分类, 即 f(x) = y 。

由于特征的多样性, 我们在使用基于特征向量的方法解决关系抽取问题时就具有很大的灵活性, 可以使用多种丰富的词法, 句法和语义特征, 我们主要归纳了以下几类特征:

(1) 词特征

这类特征根据词与实体的相对位置关系分成四类: (a) 两个实体存在所包含的词; (b) 在两个实体存在之间的词; (c) 在 M1 前面的词; (d) 在 M2 后面的词。

(2) 实体类型特征

这类特征指实体类型, 如: 人物(PERSON) , 组织(ORGANIZATION) , 地点(LOCATION) 等。

(3) 实体的具体存在类型特征

这类特征指示了同一实体的不同存在, 主要有: 唯一具体名字, 类别名称, 代词。

(4) 实例整体特征

这类特征主要指示一些全局性的特征, 如: 这个实例一共有多少个词, 两个实体间有多少个词, 以及两个实体 mention 之间具体的位置关系, 是否包含等等。

(5) 基本短语特征

由于词特征往往比较稀疏, 而短语特征有可能更好的反映实例的特征信息, 因此有的方法也加入短语特征。

(6) 依存树和句法树特征

这类特征主要是利用一些树中的路径来分析, 如: 依存树中从 M1 到 M2 的最短依存路径, 句法树从 M1 到 M2 的最短路径等等。

(7) 与其他语义资源相关特征

这类特征利用一些语义辞典来帮助分析, 如: Word Net

下面我们会用具体实验来验证这些特征在关系抽取问题中的有效性。

4 实验结果及分析

我们的实验使用的语料是由 LDC 提供的 ACE 2005 训练测试集合。ACE 2005 的语料主要来源于新闻和广播稿。ACE 2005 的 RDR 中包括三种语言的语料: 阿拉伯语, 英语和中文。我们使用其中的英文语料。由于语料是以篇章为单位的, 我们首先对其进行预处理, 以句子为单位进行切分, 提取出句子中的所有实体, 两两配对构成一个关系实例, 对于已经标注的关系实例, 我们把它归入相应的类别, 对于没有标注的关系实例, 我们归入一个无关系(non-relation) 的类别。我们抽取的训练集有 97019 个候选关系实例, 其中 8687 个是正例。测试集有 19895 个候选关系实例, 其中 1963 个是正例。对每个关系实例我们按照前面所列的特征进行特征抽取, 需要句法分析和语言分析的我们就使用对应的句法分析器(Collins Parser) 和语义资源(Word Net) 来处理。最后将每个关系实例转化为了一个带有类别标记的特征向量。

分类器我们使用开源的 lib-svm, 因为我们的实验中存在多个类别, 而这个分类器直接支持多类别问题的分类。

我们使用信息检索中的评测方法来对我们的实验结果进行评价。其主要包括了准确率(Precision) , 召回率(Recall) , F1 值三个评价量。

我们分别做了 3 个实验, 后两个实验均在前一个实验的基础上进行改进, 加入更多的特征, 并对它们进行比较以此证明我们选取的特征的有效性。第一个实验仅使用了词特征进行学习, 第二个实验在第一个实验的基础上加入了实体类型(entity type) 和实体存在类型(mention level) 这两类特征, 第三个实验在第二个实验的基础上再加入句法分析和语义信息。这三次实验的结果和比较如表 1。

表 1 实验结果及比较

	Precision	Recall	F1
词特征	0. 672	0. 268	0. 383
+ 实体类型特征	0. 669	0. 367	0. 473
+ 句法语义特征	0. 623	0. 498	0. 554

通过表 1 的比较我们可以看出我们每次加入的特征都起到了改进系统的作用, 同时我们可以发现加入更多的特征使得召回率大幅提升同时准确率在缓慢下降, 这说明特征的增加实际上使得系统判断为存在关系的实例数目在不断增加。

5 结语

本文介绍了当前信息抽取领域中热门的实体关系抽取工作, 着重介绍了其中的基于特征向量的关系抽取方法, 并结合 ACE 的语料进行了相关实验, 验证了组合特征在关系抽取问题中的有效性。

对于基于特征向量的关系抽取方法, 其核心就是要找到有效的特征来描述关系实例, 我们所列举的特征也只是其中的一部分, 对关系实例中特征的挖掘工作还要继续进行。同时我们在不断寻找特征的过程中, 也应该对这些大量的特征进行一些比较和筛选, 找到最能表达原有关系实例的特征集合。

参考文献

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

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

[3] Collins M. and Duffy N. 2001. Convolution Kernels for Natural Language. NIPS- 200.

[4] Dmitry Zelenko, Chinatsu Aone, and Anthony Rich-ardella. 2003. Kernel methods for relation extraction. Journal of Machine Learning Research, 3: 1083- 1106.