文章编号: 1000-5641(2018)03-0055-12

面向企业知识图谱构建的中文实体关系抽取

孙 晨, 付英男, 程文亮, 钱卫宁

(华东师范大学 数据科学与工程学院, 上海 200062)

摘要: 企业知识图谱是针对金融领域为描述企业间商业往来关系而构建的一类垂直领域知识库. 尽管垂直领域知识图谱在领域覆盖的广度上不如开放知识图谱, 但是它对知识准确率的要求却远远高于开放知识图谱, 因此虽然近些年开放知识图谱取得了很大的进展, 但在垂直领域中却并未得到深入应用, 尤其是商业领域, 其对企业知识图谱提出了很大的需求. 针对企业知识图谱目前在关系抽取效果上的局限性, 在分析了实体关系抽取研究现状的基础上, 提出了一种基于分类的中文实体关系抽取方法. 该方法使用最大熵模型, 通过对上市公司公报数据进行实验分析, 从而寻找到该关系抽取的最优特征模板, 并使在企业公报这一数据集上的准确率普遍达到85%以上.

关键词: 企业知识图谱; 实体关系抽取; 最大熵模型

中图分类号: TP391 **文献标志码:** A **DOI:** 10.3969/j.issn.1000-5641.2018.03.007

Chinese named entity relation extraction for enterprise knowledge graph construction

SUN Chen, FU Ying-nan, CHENG Wen-liang, QIAN Wei-ning

(*School of Data Science and Engineering, East China Normal University, Shanghai* 200062, *China*)

Abstract: The enterprise knowledge graph is a kind of domain knowledge base for the financial field to describe business relationships between enterprises. Although the domain knowledge graph is not broadly covered in the field, the precision of the knowledge is better than with an open knowledge graph. Despite the fact that open knowledge graphs have made significant advancements in recent years, vertical fields-especially business-have not seen in-depth applications in practice; this has resulted in significant demands on the enterprise knowledge graph. This paper proposes a Chinese entity relation extraction method based on classification for the limitation of extraction results. In this method, the maximum entropy model is used to analyze the data of selected companies' announcements to determine the optimal feature template. The results show that accuracy rates reach over

收稿日期: 2017-08-19

基金项目: 国家重点研发计划(2016YFB1000905); 国家自然科学基金广东省联合重点项目(U1401256); 国家自然科学基金(61672234, 61402177); 华东师范大学信息化软科学研究课题(41600-10201-562940/018).

第一作者: 孙 晨, 女, 硕士研究生, 研究方向为知识图谱. E-mail: 2683122260@qq.com.

通信作者: 钱卫宁, 男, 教授, 博士生导师, 研究方向为数据库科学. E-mail: wnqian@sei.ecnu.edu.cn.

85% in the enterprise bulletin data set.

Keywords: enterprise knowledge graph; named entity relation extraction; maximum entropy

0 引 言

随着数据规模的不断扩大, 人们从海量数据中迅速而准确地获取自己需要的信息变得越来越困难. 准确化、系统化、高效快捷的信息获取方式对信息的检索和查询提出了更高要求, 而知识图谱则是信息结构化、知识化展现的一种重要形式. 知识图谱本质上是由结点和边构成的一种语义网络, 其中结点代表分类 (Class)、概念 (Concept) 或命名实体 (Named Entity), 边代表概念或命名实体之间的语义关系 (Semantic Relationships). 具体来说, 知识图谱又可以分为开放知识图谱和垂直领域知识图谱[1-2], 其中开放知识图谱包含几乎所有领域的重要的概念、实体及其之间的关系, 强调知识的覆盖广度; 领域知识图谱则是基于某一个或若干个特定领域所构建的知识库, 强调知识的准确程度.

企业知识图谱是一类针对金融领域构建的垂直领域知识图谱, 在企业信用评估、商业欺诈检测、企业风险评估、企业行业背景调查等方面都有很多需求与应用. 它能够方便地获取和利用企业与企业之间的各种关联关系, 为企业之后的决策制定等提供具有实质性的参考意见, 同时也能够很好地改善传统 ERP 系统单一死板及非个性化等缺点. 企业知识图谱主要是由模式图和数据图两部分构成, 其中模式图通常采用本体编辑器或手工构建的方法自顶向下预先构建, 数据图则是在模式图的基础上利用多种抽取方法获得知识源中的实体、属性和关系并将这些抽取结果进行融合而得. 因此可以看出知识的获取是构建企业知识图谱的一个关键步骤, 其抽取结果的质量将直接影响知识图谱的内容表达.

本文以知识的获取为切入点, 在对实体识别完成的基础上, 主要进行实体间关系的抽取; 在分析了实体关系抽取研究现状的基础上, 提出了一种基于分类的中文实体关系抽取方法. 该方法主要使用最大熵模型, 并对上市公司公报数据集进行了实验. 上市公司数据集主要是来源于网易公司对外公开的财经新闻版块中上市公司的公报数据, 相较于一般的新闻媒体数据而言, 在语言表述上更为官方正式, 且语法结构较为单一, 便于我们对关系类型的定义与抽取. 通过相关实验分析, 我们发现其准确率普遍达到了 85% 以上.

1 相关工作

近些年来, 实体关系抽取的研究对象主要有以下几类: ① 从百科类站点和各种垂直站点获取的结构化数据; ② 从 Web 上的各种网页中爬取的半结构化数据 (如 HTML 表单等); ③ 从搜索日志中得到的数据以及一些非结构化文本数据. 对于结构化或半结构化数据而言, 我们只需要学习其固定结构就可有效地抽取出实体及相关关系, 但从非结构化文本中抽取实体关系则是一项非常艰难的研究任务: 计算机需要首先对文本片段进行语义分析, 才能从文本中识别多个实体之间的语义关系. 具体到研究方法上, 实体关系抽取的研究方法主要包括基于规则的方法 (Pattern Based)、基于有监督的统计学习方法 (Supervised Based)、基于半监督的统计学习方法 (Semi-Supervised Based) 和无监督的开放关系抽取 (Open Information Extraction) 方法.

1.1 基于规则的方法

基于规则的方法就是利用预先制定好的规则进行匹配, 其中匹配到的所有符合规则的信息就是我们将要抽取的关系. 由于该方法将关系抽取问题转化为寻找满足约束规则的关系匹配问题, 因此规则本身的学习和抽取则比目标关系的抽取更重要. Hearst等人[3]定义了一些常见的词汇语法模式 (又称 Hearst Pattern) 来抽取文本中的实体上下位 (isA) 关系. Wu等人[4]则以 Hearst Pattern 为抽取规则, 并用一定的概率来计算抽取关系的可信度. 尽管基于规则的方法能够得到较高的准确率, 但匹配得到的实体关系的召回率并不高.

1.2 基于有监督的统计学习方法

基于有监督的统计学习方法首先对标记数据集进行训练得到模型, 之后利用模型预测测试数据的关系类型. 整个系统中输入空间是各种语料, 输出空间是预先定义好的关系类型的集合. 根据训练数据 (即关系实例) 的表达方式又可细分为两类: 基于特征向量的学习方法和基于核函数的学习方法. 基于特征向量的学习方法主要是从关系实例中抽取分类器可接受的特征向量, 如句法特征、语义特征等. 文献 [5] 较为系统地介绍了如何把包括基本词组块在内的各类特征广泛结合起来, 并研究了各种语言特征对关系抽取性能的影响, 还特别研究了 Name List 和 Word Net 等语义信息对关系抽取的影响. 而基于核函数的学习方法则无需构造固有的特征向量空间, 该方法直接对结构树进行处理, 并计算结构树之间的相似度, 再使用相关分类器来抽取关系. Zhou等人[6]将关系实例描述成为上下文相关的句法解析树 (Context-Sensitive Structured Parse Tree, CS-SPT), 能够有效地根据句法结构来动态扩充与上下文相关的谓词部分, 并利用上下文相关的核函数方法进行计算, 在比较子树相似度的同时也将根结点的祖先结点考虑在内.

1.3 基于半监督的统计学习方法

基于有监督的统计学习方法虽然准确性高, 但是对大量的标注语料有很强的依赖性, 因此人们提出了基于半监督的统计学习方法. 其基本思想是以预先定义好的少量关系模式及关系实例作为种子, 然后扫描语料库得到所有符合种子模式的句子, 再将扫描得到的结果进行模式泛化 (Pattern Generalization), 泛化之后再次扫描语料库, 发现新的符合模式的实体关系对, 通过机器学习来不断获取新的实例. 最早的半监督关系抽取方法是 Brin 等人于 1998 年提出的 DIPRE(Dual Iterative Pattern Relation Expansion)[7]方法, 它利用实体关系 (Relation) 以及实体关系模式 (Pattern) 之间的对应关系, 以一个种子集合为出发点, 搜索网页中种子出现时的上下文, 从中提取对应模式, 之后利用这些模式再次搜索 Web 网页以发现更多的实例, 并从这些新获得的关系实例中选取新的种子集合, 迭代进行上述过程. 由于 DIPRE 方法在 Pattern 评估、迭代质量等方面具有一定的局限性, Agichtein 等人提出了 Snowball 系统[8]对其进行了改进.

1.4 无监督的开放关系抽取方法

无监督的开放关系抽取方法是一种完全无需人工干预的关系抽取算法, 即不需要任何预处理的语料支撑, 可以自动提取文本片段中所包含的实体间关系. 这一思想最早是由 Hasegawa[9]于 2004 年的 ACL 会议上提出, 之后的方法大多基于此改进而来. Hasegawa 的方法大致思想如下: 首先抽取句子中命名实体对, 删除出现频率较低的实例, 对于保留下来的同一命名实体对将其全部上下文整合在一起作为该实体对的上下文; 然后利用全联通聚类 (Complete Linkage) 方法对合并后的上下文进行聚类; 最后在同一类中找到出现频率最高的词语将其作为这类命名实体对的关系描述.

作为信息抽取领域一直以来的研究热点内容, 这部分工作在英文上已经有了巨大的突破, 但是还是存在一些需要在特定文本环境中才能表现良好的研究工作; 而在中文上的研究难度更是要远远大于英文, 因此基于中文的关系抽取工作仍然是一个不小的挑战.

2 任务描述

2.1 问题定义

命名实体关系抽取, 就是在给定一个由若干词语 (用 w 表示) 和两个实体 e_1 及 e_2 组成的句子 $\langle w_1, w_2, \cdots, w_i, e_1, w_j, \cdots, w_k, e_2, w_m, \cdots, w_n\rangle$ 中抽取出实体 e_1 和 e_2 之间的语义关系. 具体到本文的研究方法上, 我们主要是将关系抽取转变为训练分类器问题, 即给定包含实体对的上下文语言环境和具体的实体关系, 判断实体对之间是否满足该关系.

基于此, 首先需要确定待抽取的关系类型. 由于本文主要是针对上市公司的公报数据, 因此描述企业间商业往来关系的词汇相对而言比较集中而明确, 所以我们可以通过人工预先定义实体关系类型. 鉴于动词能够很好地表达实体间的关系类型[10], 因此我们通过统计每一个动词出现的频数, 按降序排序后抽取高频动词并以其为出发点来扩展定义所需的关系类型. 因为这些动词词汇普遍是领域术语, 所以我们就直接对其出现次数进行了统计. 表 1 给出了词频排序中排在前十的动词, 可以看出基本符合我们的需求, 表 2 就基于这些原始词汇列出了一小部分关系类型和描述该类型的一些其他词语. 然后我们就可以针对每种关系定义相对应的特征模板, 并标注训练数据, 进而训练模型, 得到关系抽取的分类器. 最后, 我们对该分类器进行测试分析, 并寻找最优特征模板.

表 1 动词词频前十排序

Tab. 1 Top ten list of verbs

动词	词频
持有	17 553 963
投资	14 848 084
发行	11 762 727
转让	11 015 165
简称	7 799 445
合并	7 403 418
签订	6 024 973
委托	5 246 879
购买	4 969 604
收购	4 810 237

表 2 关系类型定义

Tab. 2 Relationship type definition

关系类型	描述关系的动词集合
持有	持有, 合计持有, 应付持有, 合并持有, 转让持有, 拟转让持有,.....
投资	投资, 对外投资, 投资建设, 合作投资, 投资并购, 投资担保,.....
转让	转让, 挂牌转让, 出资转让, 限制转让, 作价转让, 优先转让,.....
合并	合并, 吸收合并, 进行合并, 购买合并, 合并收购, 收购合并,.....
收购	收买, 并购, 收购, 拟收购, 要约收购, 收购完成, 出资收购,.....

2.2 最大熵模型

我们选择最大熵 (Max Entropy) 模型[11-12]来训练数据, 主要有以下几点原因: 第一, 相比于朴素贝叶斯[13]等模型而言, 最大熵模型可以灵活地使用特征方程来表达特征之间的约束关系, 这样既便于将特征间的复杂关系转化为数学公式的形式, 又可以直接通过修改约束条件来调节模型对训练数据的拟合度以及对测试数据的适应度; 第二, 相较于非概率模型如基于转换学习

等, 最大熵模型可以利用其概率分布对多个分类结果进行评估, 从而选出最优化结果; 第三, 尽管决策树模型在自然语言处理上比较成功, 但由于决策树模型过度依赖于聚类算法而无法很准确地消除歧义问题, 因此选用最大熵模型, 可以直接地利用单词而无需担心数据碎片; 第四, 条件随机场 (Conditional Random Field) 模型[14]中虽然也会大量使用到特征方程, 但条件随机场更适合于标注问题, 因为条件随机场是对给定序列的联合概率进行建模, 所以相较之下最大熵模型更适用于解决我们的问题. 此外, 最大熵模型对于没有指定约束条件的特征之间的关系不做任何假设, 即认为所有结果等可能性发生, 这也较为符合人们正常的逻辑思维方式. 模型中将会使用到的一些数学公式符号及其定义如表 3 所示.

表 3　符号定义

Tab. 3　Symbol definition

符号	描述
X	经过分词处理后的词向量的有限集合
Y	输出状态的有限集合
x	任一词向量
y	任一输出状态
$p(y\|x)$	给定上下文 x 的情况下输出 y 的概率
$\widehat{p}(y\|x)$	给定上下文 x 条件下输出 y 的概率的经验分布
$p(x,y)$	给定上下文 x 且输出 y 的联合概率分布
$\widehat{p}(x\|y)$	给定上下文 x 且输出 y 的联合概率的经验分布
$\widehat{p}(x)$	x 在样本中出现的概率分布
$f(x,y)$	表征输入特征 x 和输出状态 y 之间关系的特征方程
$Ep\langle f_i\rangle$	特征函数关于模型 $p(y\|x)$ 的期望
$E\widehat{p}\langle f_i\rangle$	特征函数关于经验分布 $\widehat{p}(y\|x)$ 的期望
$Z(x)$	规范化因子
λ_i	特征的权值

最大熵模型的主要思想如下: 考察这样一个随机系统, 已知其输出是 y, $y\in Y$, 其中 Y 是输出状态的有限集合; 而在输出 y 时, 它会受到句子中上下文信息 x 的影响, $x\in X$, 其中 X 是经过分词处理后的词向量的有限集合. 那么我们就可以根据这些信息来构造一个统计模型来计算在给定上下文 x 的情况下输出 y 的概率 $p(y|x)$, 从而进行相关预测. 根据我们在一段时间内观察得到的大量样本数据, 可以统计出在给定上下文 x 条件下 y 的经验分布 $\widehat{p}(y|x)$, 接下来就可以利用最大熵模型来构建模型以使其无限逼近该随机系统, 即估计条件概率 $p(y|x)$ 使得它尽可能接近于 $\widehat{p}(y|x)$.

基于训练样本的各种统计信息, 我们可以定义特征方程 $f(x,y)$ 来表征输入特征 x 和输出状态 y 之间是否满足某一事实或具有某些特定关系, 如果满足条件则取 1 否则取 0. 比如, 若两个企业实体之间存在"合作"关系, 且上下文中出现了"合作"这一关键词, 则特征方程 $f(x,y)=1$. 具体的特征方程为

$$f(x,y)=\begin{cases}1, & \text{如果上下文 } x \text{ 中含有关键词"合作"\&\& } y==1,\\ 0, & \text{其余情况.}\end{cases}\tag{1}$$

这样我们就可以灵活地定义一系列表征正例和负例的特征方程.

在选定特征集合后, 我们就需要让模型去符合它, 这一拟合过程可通过约束模型 p 分配给它相应特征函数的期望值来实现. 特征函数关于模型 $p(y|x)$ 的期望值是

$$Ep\langle f_i\rangle=\sum_{x,y}p(x,y)f_i(x,y)=\sum_{x,y}\widehat{p}(x)p(y|x)f_i(x,y).\tag{2}$$

而我们也很容易从历史数据中挖掘出特征函数关于经验分布 $\widehat{p}(y|x)$ 的期望值, 即

$$E\widehat{p}\langle f_i\rangle = \sum_{x,y}\widehat{p}(x,y)f_i(x,y). \tag{3}$$

为了使模型也拥有训练数据中存在的统计现象, 我们约束上述二者相等, 可得到方程

$$E\widehat{p}\langle f_i\rangle = Ep\langle f_i\rangle. \tag{4}$$

因此 $p(y|x)$ 满足等式

$$\sum_{x,y}\widehat{p}(x,y)f_i(x,y) = \sum_{x,y}\widehat{p}(x)p(y|x)f_i(x,y). \tag{5}$$

在满足上述条件的前提下对所有情况, 我们最合理的考虑就是认为其等可能发生, 换言之使均匀度最大. 在数学定义中, 条件分布 $p(y|x)$ 的均匀度就是条件熵. 因此我们使用最大熵模型: 在满足上述约束的条件下, 选择模型使得条件熵最大, 即

$$H(p) = -\sum_{x,y}\widehat{p}(x)p(y|x)\log p(y|x). \tag{6}$$

构建最大熵模型, 就是求解如下带约束条件的优化问题.

$$\begin{aligned}&\max_p H(p) = -\sum_{x,y}\widehat{p}(x)p(y|x)\log p(y|x),\\ &\text{使得 } E\widehat{p}\langle f_i\rangle = Ep\langle f_i\rangle, \quad i = 1,2,3,\cdots,n,\\ &\sum_y p(y|x) = 1.\end{aligned}$$

通过拉格朗日对偶求上述带约束的目标函数的最优解, 得到最优解的表达式

$$p(y|x) = \frac{1}{z(x)}\exp\Big[\sum_{i=1}^{n}\lambda_i f_i(x,y)\Big], \tag{7}$$

其中

$$Z(x) = \sum_y \exp\Big[\sum_{i=1}^{n}\lambda_i f_i(x,y)\Big] \tag{8}$$

是规范化因子, 确保 $\sum\limits_y p(y|x) = 1$, $f_i(x,y)$ 是特征函数, λ_i 表示特征的权值. 接下来模型就要去学习参数 λ_i.

根据最大似然估计法, 估计模型的似然函数为

$$L(p) = \log\prod_{x,y}p(y|x)^{\widehat{p}(x,y)} = \sum_{x,y}\widehat{p}(x,y)\log(y|x). \tag{9}$$

因为直接使用最大似然法来估计一个指数线性函数是很难的, 所以就提出了广义迭代尺度法 (Generalized Iterative Scaling, GIS)[15]、改进的迭代尺度算法 (Improved Iterative Scaling, IIS)[16]等来解决该参数学习问题.

3　特征选取

因为特征的好坏经常能够直接影响到模型的性能, 所以特征的选取就变得尤为重要. 在自然语言处理中, 特征主要是从自由文本及其语法结构中抽取而来, 包括各种字面特征以及经过语法分析后的结构化特征. 因此本文就从构成文章的小粒度词汇出发, 同时深入词汇之间的句法、语义关系, 以期得到较为全面而有效的特征空间.

3.1　基本词汇特征

在进行实体关系的抽取时, 给定的上下文中首先要出现能描述该关系类型的词语, 比如要抽取"合并"关系, 根据表2中关系类型的定义, 句子上下文中要出现类似"合并"、"吸收合并"、"购买合并"这样的动词词汇, 本文中将这类动词称为关键词. 除了关键词以外, 实体词汇也是关系抽取中的一类重要对象, 因为它说明了关系的施事者与受事者, 有助于我们理解关系的构成. 因此我们就从这两类中心词汇出发, 结合在其一定范围内的其他词汇对关系类型的补充描述, 总结出词汇特征的主要构成: 首先是关键词和实体本身的词汇及词性; 其次是关键词和实体左右一定窗口范围内的词汇和词性, 本文定义的窗口大小为2; 再次是关键词和实体的长度; 最后是关键词在句子中的位置以及和实体间的相对顺序及位置间隔. 具体词汇特征模板如表4所示.

表 4　词汇特征模板

Tab. 4　Vocabulary feature template

特征描述	特征表示
关键词、实体上下文1-Gram词汇特征	w_{i-2},w_{i-1}, w_i, w_{i+1}, w_{i+2}
关键词、实体上下文1-Gram词性特征	t_{i-2}, t_{i-1}, t_i, t_{i+1}, t_{i+2}
关键词、实体上下文2-Gram词汇特征	$w_{i-2}w_{i-1}$, $w_{i-1}w_i$, w_iw_{i+1}, $w_{i+1}w_{i+2}$
关键词、实体上下文2-Gram词性特征	$t_{i-2}t_{i-1}$, $t_{i-1}t_i$, t_it_{i+1}, $t_{i+1}t_{i+2}$
关键词、实体长度特征	$\mathrm{len_{key}}$, len_e
关键词、实体位置特征	p_{key}, $(p_{\mathrm{key}}-p_{e1})$, $(p_{e2}-p_{\mathrm{key}})$

因为本文主要针对的是企业之间的关系抽取, 而根据大量训练数据的统计结果发现, 对于每一类关系, 某些特定词语的出现会对该关系是否成立起重大影响. 以"合并"关系为例, 它一般是指一家企业与另一家企业进行合并, 所以若在中心词"合并"的左侧有两个相距不远的企业实体, 那么该句子中的"合并"关系很有可能成立. 再比如对于从属关系中的关键词"母公司", 它往往指的是一家公司是另一家公司的母公司或者一家公司的母公司是某某公司, 基于此模式一般在以"母公司"为中心的上下文中, 如果在关键词"母公司"左侧有两个企业实体或者在距离关键词"母公司"不远的左右两侧都有一个被命名实体识别为NTC的实体且右侧实体之前的词汇为"是"或"为"之类, 那么该句子中的"母公司"关系很可能成立, 而且实体离"母公司"越近可能性越大. 从上述例子当中我们也能看出不同的关系类型需要定义不同的具体词汇搭配特征. 表5就以"母公司"这一关系为例, 列出了一系列"母公司"关系所具备的特征.

表 5　"母公司关系"的特定词汇特征

Tab. 5　Specific lexical features of relationship between the parent companies

特征名称	特征描述
$t-i$=:nt	关键词前面第 i 个词语的词性包含 nt
$t+j$=:nt	关键词后面第 j 个词语的词性包含 nt
$t-1$=uj	关键词前面一个词语的词性等于 uj
$w-1$=的	关键词前面一个词语的词汇等于"的"
$t+1$=n	关键词后面一个词语的词性等于 n
$w+1$=名称	关键词后面一个词语的词汇等于"名称"
$t-i$=:v 和$w-i$=:是	关键词前面第 i 个词语的词性包含 v 并且词汇包含"是"

3.2 全局句法特征

鉴于词汇特征主要是对给定的句子进行局部的特征抽取, 没有对句子的整体架构进行分析, 因此引入句法特征, 综合全局来进行相关特征的抽取. 句法分析主要包含成分句法分析、依存句法分析等, 一般通过句法结构框架树的形式展示出结果. 成分句法分析主要是识别出句子中的短语句法结构, 比如介词短语、名词短语等, 然后在此基础上确定主次成分, 如主语、谓语、宾语一般充当句子的主要成分, 而定语、状语、补语则是次要的修饰成分.

图 1 为句子"1984 年青岛化肥厂更名为青岛碱厂, 1993 年由国家有关部委核定为大型一档企业"的成分句法分析树. 从图中我们可以发现这一长句 (IP) 主要由两个短句 (IP) 构成, 而短句又都由名词短语 (NP) 和动词短语 (VP) 构成, 其中名词短语和动词短语又可以细分为更多更具体的结构. 具体的结构名称及含义如表 6 所示. 此外, 我们还可以从中获取一组语义关系三元组即(青岛化肥厂, 更名, 青岛碱厂)以及实体和关键词间的路径状况.

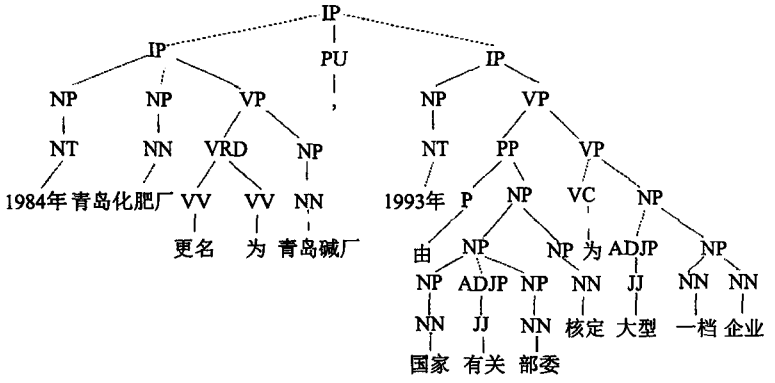

图 1 成分句法分析样例

Fig. 1 Sample composition of syntactic analysis

表 6 句子成分

Tab. 6 Sentence composition

符号	句子成分描述	符号	句子成分描述
ADJP	形容词短语	PP	介词短语
IP	句子	PU	标点符号
JJ	名词修饰语	VC	系动词
NN	普通名词	VP	动词短语
NP	名词短语	VRD	动词结果复合
NT	时间名词	VV	普通动词
P	介词		

依存句法分析是通过分析给定句子中各个成分之间具体的依存关系来显示其句法结构, 强调句子中的关键词是支配其他成分的核心成分, 而其本身却不受任何别的成分的支配, 所有受支配者均以某种依存关系从属于该支配者[17]. 图 2 就展示了上述例句的依存句法分析树, 从图中可以看出, 在第一个短句中以关键词"更名"为句子的核心, 并由它出发指向句子的其他部分, 包括主语部分"青岛化肥厂"和宾语"青岛碱厂"以及其他的修饰部分 (MOD), 比如"1984 年"这一时间状语. 因此本例中的"青岛化肥厂 $\longrightarrow$SUB$\longrightarrow$更名$\longleftarrow$OBJ$\longleftarrow$青岛碱厂"就可以作为依存句法分析的一条正例而加入到句法特征函数中, 其中具体的实体名称可以用 NN 来统一代指.

总结上述分析, 我们将会从以下几个方面来抽取句法特征: 首先是连接关键词的依存关系类型及依存关系成分的词性; 其次是实体与关键词以及实体与实体之间的依存路径和成分路径;

最后是实体与关键词及实体与实体之间的间距, 比如上例中"青岛化肥厂"与"青岛碱厂"之间的距离为 3.

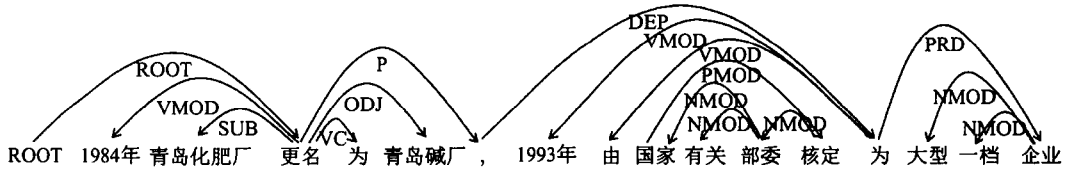

图 2　依存句法分析样例

Fig. 2　Sample of interdependence syntactic analysis

4　实验分析

这一部分主要是通过具体实验进行相关分析. 对于基于统计学习方法的关系抽取, 首先需要标注一定的训练数据, 然后在此训练集上选取特征, 其中不同的特征模板及其彼此间的组合对于模型的影响也有所不同. 所以接下来我们会主要介绍模型的评估方法以及不同特征的组合对模型的影响情况.

4.1　实验设置

本文针对不同类型的语义实体关系平均手工标注了约 2 000 条左右正例和 3 500 条左右负例, 从中随机选择 80% 作为训练语料, 20% 作为测试语料进行了实验. 同时我们采用了张乐博士所开发的最大熵工具包, 通过训练最大熵模型, 以期找到最优特征模板. 获取实体关系特征模板的核心算法见算法 1.

算法 1　获取特征模板

输入: 训练数据的集合, Set⟨sample⟩.

输出: 训练数据特征向量的集合, List⟨featureVector⟩

```
put all the entities into ANSJ.
Result = {}.
for each sample in Set⟨sample⟩ do
    get sample label and put the label into featureVector;
    get two entities and keyword then get unigram word of them in the sliding window for 2;
    for each unigram word do
      put it into featureVector and label it as 1;
    end for
    get unigram position feature of the entities and keyword in the sliding window for 2;
    get bigram word and position feature of the entities and keyword also in the sliding window for 2;
    then get the length of the entities and keyword;
    get the distance between entities and keyword;
    put all the features into featureVector like the unigram features;
    add all featureVectors into Result;
end for
return Result;
```

由于本文将关系抽取视为二分类问题, 即属于某一类的样本就被定义为"正例", 反之为"负例", 所以这里也就采用常规的准确率[18] (Accuracy)、查准率[18] (Precision)、召回率[18]

(Recall) 以及它们的平衡指标 F1-Measure[18]等这些性能评估标准. 需要注意的是当正例和负例两类数据相对均衡时, 上述指标会有较好的表现效果. 具体公式如下.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \times 100\%, \tag{10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\%, \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\%, \tag{12}$$

$$\text{F1-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \times 100\%, \tag{13}$$

其中 TP(True Positive) 表示正例被模型正确预测为正例的数量, FN(False Negative) 表示正例被错误预测为负例的数量, FP(False Positive) 表示负例被错误预测为正例的数量, TN(True Negative) 表示负例被正确预测为负例的数量. 式 (10) 表示准确率, 即分类器将正例正确预测为正例、将负例正确预测为负例的比例, 是对整个训练集的判定能力; 式 (11) 表示查准率, 即在模型预测为正例的所有样本中真正正例的比例, 表征分类器识别正例的正确率; 式 (12) 表示召回率, 即被模型正确分类为正例的个数占实际总的正例个数的比例, 表征分类器识别正例的覆盖率; 式 (13) 表示 F1-Measure, 是查准率和召回率的调和平均值, 是对二者的一个综合评估, 当该值较高时说明模型比较有效.

4.2 实验结果分析

为了找到该关系抽取分类器中的最优特征, 本文对各类特征及其组合的效果进行了分析对比, 共有如下 7 组实验: ① 一般词汇特征; ② 特定词汇特征; ③ 全局句法特征; ④ 两种词汇特征的组合; ⑤ 一般词汇特征和全局句法特征的组合; ⑥ 特定词汇特征和全局句法特征的组合; ⑦ 3 种特征的组合. 以“更名关系”为例, 具体实验结果如图 3 至图 6 所示.

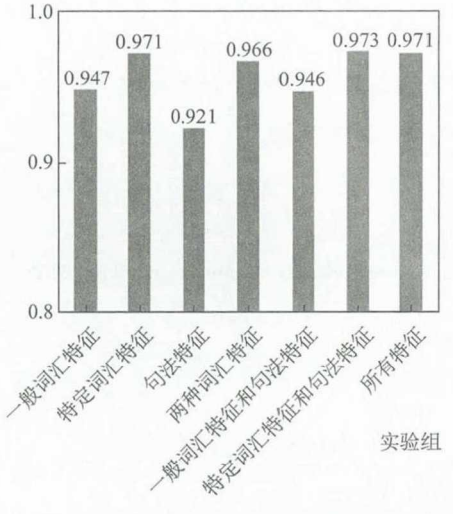

图 3 查准率

Fig.3 Precision assessment

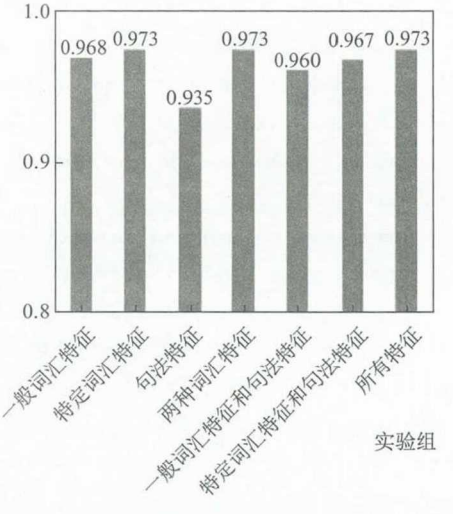

图 4 召回率

Fig.4 Recall assessment

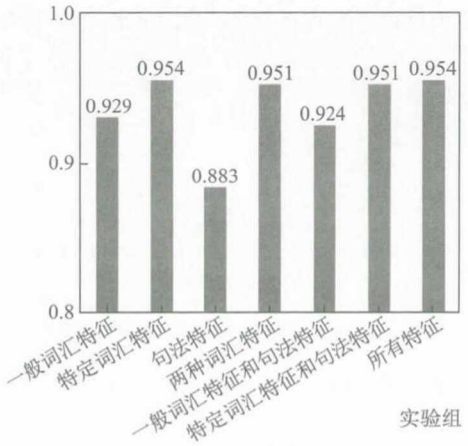

图 5　准确率

Fig.5　Accuracy assessment

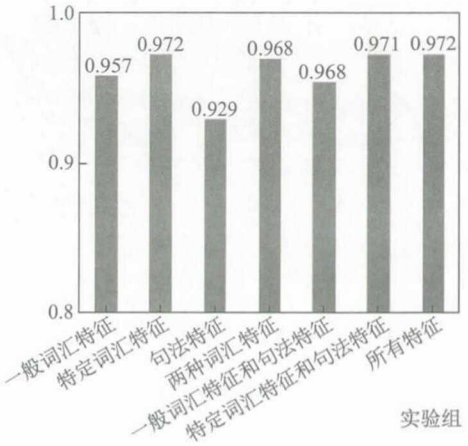

图 6　F1-Measure

Fig.6　F1-Measure assessment

首先对比前3组单一特征实验, 可以发现按效果优劣排序, 依次是特定词汇特征、一般词汇特征、全局句法特征. 特定词汇特征之所以在3者中拥有最好的区分效果, 主要是因为该类特征是针对每一类关系的细节化语法特征, 且鉴于更名关系本身的一些特性, 所以局部的细节特征更便于提取实体间的语义关系. 同时由于最大熵模型以特征函数为约束条件, 而特定词汇特征所描述的特征约束考虑的细节较为全面准确, 所以使得模型的表现效果相对而言也较好. 而一般词汇特征尽管粒度更小, 但由于考虑到更多的语法现象所以反而会使数据略有过拟合的现象. 对于句法特征而言, 由于它只是获取句法依存关系、依存路径等全局特征, 相比于前两者, 其粒度最为抽象, 因而导致模型欠拟合, 表现效果最差.

接着我们对比分析后4组组合特征实验, 可以发现一般词汇特征、特定词汇特征和全局句法特征3种组合即最后一组实验的表现效果最好, 然后依次是特定词汇特征和句法特征的组合、特定词汇特征和一般词汇特征的组合以及一般词汇特征和句法特征的组合. 由于一般词汇特征是从最根本的关键词语、实体词语等出发进行研究, 句法特征则是立足于词汇之间的语法结构, 而特定词汇特征主要是对词汇之间的关联搭配进行相关研究, 所以将3个特征组合起来实际上是将对句子3个方面的研究融合在了一起, 非常全面, 也使得模型的表现效果相对最优. 而一般词汇特征和句法特征的组合表现最差, 主要是因为其特征内容不够具有针对性, 特征范围选择上过宽或过窄, 没有从一个合适的范围去描述句子的特征, 使得特征约束不够, 模型所能学习到的知识较少, 表现较差.

最后对比分析前3组单一特征实验和后4组组合特征实验, 可以发现, 特定词汇特征的表现效果已经达到了较优的程度, 即使添加上另外两个特征也基本上没有太大的提升, 相反有时还会减弱其表现性能. 所以针对更名关系我们只需要构建它的特定词汇特征即可.

从整体分析来看, 因为最大熵模型没有对特征之间做很强的独立性假设, 所以其充分考虑到了许多词汇之间的多重关联关系, 并且将这些未知的多重关联关系合理地假设为均匀分布, 从而使得整体的表现效果都较为不错.

5　结　　论

本文以最大熵模型为例, 从机器学习的角度对实体关系抽取进行了研究, 主要从文本内容、具体词汇搭配以及句法结构3个方面定义了一般词汇特征、特定词汇特征和全局句法特征3类模板, 并通过实验发现特定词汇特征本身就可以使模型的表现效果达到最优, 其准确率基本可达到97%. 基于此, 我们从109万篇上市公司的公报数据和250万篇新闻数据中抽取了超

过5万以上的命名实体并提取出了14万以上的实体关系, 构建出了较为完整的企业知识图谱. 这对于构建垂直领域知识图谱有一定的实践意义, 同时对于开放知识图谱的构建也具有参考价值.

[参 考 文 献]

[1] PUJARA J, MIAO H, GETOOR L, et al. Knowledge graph identification [C]//International Semantic Web Conference. New York: Springer-Verlag, Inc, 2013: 542-557.

[2] DESHPANDE O, LAMBA D S, TOURN M, et al. Building, maintaining, and using knowledge bases: A report from the trenches [C]//ACM SIGMOD International Conference on Management of Data. ACM, 2013: 1209-1220.

[3] HEARST M A. Automatic acquisition of hyponyms from large text corpora [C]//Proceeding of the 14th Conference on Computational Linguistics. 1992: 539-545.

[4] WU W T, LI H S, WANG H X, et al. Probase: A probabilistic taxonomy for text understanding [C]//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012: 481-492.

[5] ZHOU G D, SU J, ZHANG J, et al. Exploring various knowledge in relation extraction [C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 2005: 427-434.

[6] ZHOU G D, ZHANG M, JI D H, et al. Tree kernel-based relation extraction with context-sensitive structured parse Tree information [C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. DBLP, 2007: 728-736.

[7] BRIN S. Extracting patterns and relations from the World Wide Web [C]//WebDB'98 Selected Papers from the International Workshop on the World Wide Web and Databases. Berlin: Springer, 1998: 172-183.

[8] AGICHTEIN E, GRAVANO L. Snowball: Extracting relations from large plain-text collections [C]//ACM Conference on Digital Libraries. ACM, 2000: 85-94.

[9] HASEGAWA T, SEKINE S, GRISHMAN R. Discovering relations among named entities from large corpora [C]//Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004:415.

[10] 郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取 [J]. 中文信息学报, 2014, 28(6): 183-189.

[11] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [C]//Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2004: Article No 22.

[12] RATNAPARKHI A. Maximum entropy models for natural language ambiguity resolution [D]. Pennsylvania: University of Pennsylvania, 1998.

[13] 李丹. 基于朴素贝叶斯方法的中文文本分类研究 [D]. 石家庄: 河北大学, 2011.

[14] 薛俊欣. 条件随机场模型研究及应用 [D]. 济南: 山东大学, 2014.

[15] DARROCH J N, RATCLIFF D. Generalized iterative scaling for log-linear models [J]. Annals of Mathematical Statistics, 1972, 43(5): 1470-1480.

[16] BERGER A. The improved iterative scaling algorithm: A gentle introduction [R/OL]. (1997-12-12)[2017-05-19]. http://www.doc88.com/p-1806889293798.html.

[17] 胡宝顺, 王大玲, 于戈, 等. 基于句法结构特征分析及分类技术的答案提取算法 [J]. 计算机学报, 2008, 31(4): 662-676.

[18] OLSON D L, DELEN D. Advanced Data Mining Techniques [M]. Berlin: Springer, 2008.

(责任编辑: 李 艺)