中文金融新闻中公司名的识别

王 宁[1] 葛瑞芳[1] 苑春法[1] 黄锦辉[2] 李文捷[3]

(1. 智能技术与系统国家重点实验室 清华大学计算机科学与技术系 北京 100084
2. 香港中文大学系统工程与工程管理系 3. 香港理工大学电子计算学系)

摘要: 在金融领域信息抽取中, 公司名扮演着非常重要的角色; 因此如何正确识别文本中出现的公司名是一个非常重要的研究课题。在对金融新闻文本进行了深入地分析和研究的基础上, 总结出了公司名的结构特征及其上下文信息, 建立了六个用于识别公司名的知识库, 并提出了一个基于两次扫描过程的识别策略。初步实验结果表明, 在封闭测试中实验系统公司名识别的精确率可以达到 97.3%, 召回率可达 89.3%; 在开放测试中精确率可以达到 62.8%, 召回率可达 62.1%。

关键词: 公司名; 金融领域; 专名识别; 信息抽取

中图分类号: TP391.4

Company Name Identification in Chinese Financial Domain

WANG Ning[1] GE Rui-fang[1] YUAN Chun-fa[1] K. F. Wong[2] LI Wen-jie[3]

(1. State Key Laboratory of Intelligent Technology and System
Dept. of Computer Science & Technology Tsinghua University Beijing 100084
2. Dept. of System Engineering & Engineering Management The Chinese University of Hong Kong Hong Kong
3. Department of Computing The Hong Kong Polytechnic University Hung Hom Hong Kong)

Abstract: Identifying company names in running texts plays a significant role in financial information extraction. Based on the thoroughly investigations of financial articles, the relevant structural features and contextual constraints were obtained. In this paper, a company name identification system is proposed, which is built on the six knowledge-bases and a twice-scan method. The experiment achieved 97.3% precision and 89.3% recall respectively by close test, and 62.8% precision and 62.1% recall respectively by open test.

Keywords: company name; financial domain; named entity identification; information extraction

一、引言

信息抽取(Information Extraction, IE)目前已经成为继机器翻译(Machine Translation, MT)和信息检索(Information Retrieval, IR)之后得到各国政府和企业界普遍关注的一个重要的应用领域[6]。

专名识别是信息抽取系统底层的预处理子系统之一, 它的任务是将文章中出现的专有名词(Named Entity, NE)如时间、日期、公司名称等识别出来[5]。专名识别的完成水平直接关系到信息抽取的质量。美国国防部资助的系列会议 MUC(Message Understanding Conference)是

收稿日期: 2001-11-20
本文得到国家自然科学基金(69975008)和国家重点基础研究 973(G1998030507)项目支持
作者王宁, 男, 1977 年生, 硕士研究生, 主要研究方向为自然语言处理. 葛瑞芳, 女, 1976 年生, 硕士研究生, 主要研究方向为自然语言处理. 苑春法, 男, 1946 年生, 教授, 主要研究方向为自然语言处理.

在信息抽取领域中影响最大的会议。在 MUC 中,专名的识别主要分为三个子任务:名字的识别(ENAMEX),包括人名、地名、机构名;时间的识别(TIMEX),包括对时间短语如日期、时间等的识别;数字的识别(NUMEX),包括对金钱数量和百分比数量的识别等。和第一个子任务相比,后面两个子任务几乎完全可以依靠几种模式匹配完成,要简单的多。因此名字的识别(ENAMEX)是专名识别研究的重点。

关于名字的识别(ENAMEX)国内外都已经有了大量的研究工作,尤其是在人名和地名的识别方面,而在机构名的识别方面相对比较少,涉及到中文机构名识别的更少。机构名包括学校、公司、医院、研究所和政府机关等,在金融领域中,机构名主要是公司名。公司名在金融领域中扮演着非常重要的角色,本文所讨论的是广义的公司名,包括公司、企业、银行、厂矿、股票名称,等等。目前国内外重要刊物上还没有关于中文金融领域中公司名识别研究工作的报导。

香港理工大学[1]对中文机构名称尤其是中文高校名称的组成和特征进行了深入的分析,并采用基于规则的方法对高校名称进行识别,取得了很好的效果,在六百多万字的测试集上准确率和召回率分别为 97.3%和 96.9%。

文献[2]采用基于规则的方法建立了一个专名识别系统 NTUNLPL,并参加了 MUC-7 专名任务的评比,F1 测试值为 79.61%。其中机构名识别部分的准确率为 85%,召回率为 78%。

Georgetown 大学指出了机构名识别是 ENAMEX 的识别中最困难的部分[3],但是没有对机构名进行专门处理,只是把它当作 NE 的七种类型之一通过模式匹配进行识别,对于公司名更是没有任何专门处理,因此效果不太理想。专名识别的准确率和召回率在含有 1117 个 NE 的测试集上为 53%和 46%,在含有 254 个 NE 的测试集上为 29%和 17%。

Keh-Jiann Chen 等从未登录词处理的角度对中文机构名的结构特征进行了分析,主要依靠规则的方法对中文机构名称进行识别,在 31787 个新闻文本的测试集上,一个比较好的结果是准确率和召回率分别为 61.79%和 54.50%[4]。所采用的新闻文本主要属于金融领域,在识别工作中也借助了一些公司名的结构特点,但并没有针对金融领域的特点进行全面深入的分析。

本文在从专名识别处理的角度进行研究工作的同时,结合了前人在机构名识别研究上的可取之处,充分利用了金融领域的特征,专门针对公司名的识别问题进行研究。在识别策略上综合考虑了公司名的结构特征和文本上下文信息,建立了六个用于识别的知识库,并提出了一个基于两次扫描过程的识别策略,实验结果是令人满意的。

文章其余部分的结构如下:第二部分介绍公司名在金融领域中的特征;第三部分介绍公司名识别知识库;第四部分介绍识别策略;第五部分介绍实验结果以及对实验结果的分析;第六部分是结束语。

二、公司名特征分析

公司名属于"定语+名词性中心语"型的名词短语,简称定名型短语,从宏观看来是一种偏正式复合名词,其结构为 $X^{+}Y$,其中'X'和'Y'表示词,X^{+}表示 X 元素可以出现一次或多次。公司名的中心语主要集中在"公司"、"集团"等有限的一些名词上。这对我们识别公司名的右边界起到了非常大的作用。

有不少公司名是以地名或人名开头,这对我们识别公司名的左边界是有一定的作用的。在研究了大量的真实文本后,我们发现在公司名中,有些词和有些词性的词是明显不会作为公司名组成部分的。由于公司名的识别过程是从右边界向前扫描的,因此这类的词或者词性恰

巧可以作为我们初步判定左边界的依据。同时，还可以通过考察文本中公司名出现位置的前邻近词或者在文本中进行局部统计来辅助判断公司名的左边界。

公司名的出现情况有两种：全称，简称。

所谓公司名全称，就是公司名的最正式的叫法，一般由地名、公司名关键字、公司类型名和公司名后缀组成。例如“北京新唐装饰工程有限公司”，包括了地名“北京”，公司名关键字“新唐”，公司类型名“装饰”、“工程”、“有限”，公司名后缀“公司”。

公司名的简称是对公司名全称缩略的叫法，对于含有关键字的公司名全称，公司名的简称一般都包含公司名关键字，而地名、类型名和后缀都是可选的部分，可以有，也可以没有；对于含有多个类型名的公司名，简称中可以包含任意一个类型名，也可以含有多个类型名。例如“深圳市赛百诺基因技术有限公司”，既可以简称为“赛百诺公司”，也可以叫做“赛百诺基因公司”。

公司名的简称大体上可分为四种类型，如下表：

1. 缩写	武汉钢铁	武钢
	三明啤酒厂	三啤
2. 连续简写	联想集团	联想
	上海紫江企业集团股份有限公司	紫江企业
	江阴长江科技投资有限公司	长江科技
	西北轴承股份有限公司	西北轴承
	昆明云内动力股份有限公司	云内动力
3. 不连续简写	北京新唐装饰工程有限公司	新唐公司
	福建省绿得罐头饮料有限公司	绿得公司
	美国福特汽车公司	福特公司
4. 简写缩写混合	东风汽车电子仪表股份有限公司	东风电仪

由此也可以看出，公司名关键字的识别，对于公司名简称的识别具有很重要的意义。

汉语类似于日语，是一种缺乏形态标记的语言，如果采用英文专名识别的方法，主要依靠机器学习或者统计的方法，效果必然是不很理想。因此公司名的识别应主要依靠公司名的组成规则进行判断（例如：“公司名＝地名＋{公司名关键字}*＋{公司类型名}*＋公司名后缀”等等，这里上标*表示集合中的元素为可选项），结合统计的方法，同时也可借助一些上下文信息。

我们主要借助了两种上下文信息：

1. 边界词：我们认为“兼并”、“收购”之类领域特有动词的后面，以及“的股权”之类的领域常用字串前面，都很有可能是公司名。

2. 并列关系词：在并列关系下，如“和”、“与”等词的一侧如果是公司名，则另一侧很有可能是公司名。

三、公司名识别知识库

本文使用的语料库由来自网上的近期金融新闻构成，共 1，336 篇，约有 8M 的真实文本。首先，我们对金融领域语料库进行了分词、词性标注和公司名标注，形成了公司名的标准语料库。其次，我们对北京和上海的共计 87，500 条公司名进行了分词、词性标注，得到了公司名识别必要的资源。在这些资源的基础上，利用机器统计和人工辅助相结合的方法，建立了如下知识库，用来指导公司名的识别。

3.1 公司名后缀库

在我们的系统中，对公司名的识别首先从确定公司名右边界开始，而例如“公司”、“银行”、“集团”、“企业”之类的词可以提供准确的公司名右边界信息，因此将它们集中起来，建立公司

名后缀库，作为识别的触发条件。

3.2 公司类型名库

公司类型名包括“投资”、“开发”、“有限责任”等附加在公司名后缀前的词。这些词不象公司名后缀那样有限，又不象公司名关键字那样形式多样，因此专门建立公司类型名库，辅助系统进行识别。

3.3 公司名禁止词性库

有些词性的词是明显不会作为公司名组成部分的，而这类的词或者词性恰巧可以作为我们初步判定左边界的依据。为此，我们建立“公司名禁止词性库”，在识别过程中，碰到库中的词性则中断公司名的识别，认为当前识别对象不是公司名的组成部分。通过深入分析，我们认为这些词性主要集中在虚词、连接词和动词，如：vgs(动词带小句宾语)，pzai(在)，c(连词)等。

3.4 公司名禁止词库

有些词的词性允许在公司名中出现，但是该词并不允许在公司名中出现，主要是一些形容词、名词如“优秀”、“股权”、“股东”等。因此我们把它们集中起来建立“公司名禁止词库”，在识别过程中，碰到库中的词则中断公司名的识别，认为当前识别对象不是公司名成分。

3.5 公司名完全禁止库

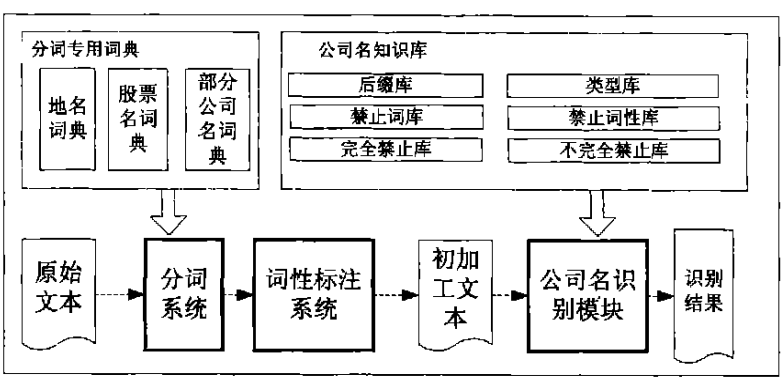
有一种情况，就是当从公司名右边界开始向左搜索到某个位置时，在当前位置就可以断定这不是公司名，从而停止这个串的搜索。例如上市公司，其他公司，第一大公司，两家公司，三个企业，科技类企业，高新技术企业，国有(大中型)企业，国有商业银行，蓝筹绩优公司，子公司，破产企业，等等。“公司名完全禁止库”收录了这些完全禁止词串。

3.6 公司名不完全禁止库

还有一种情况，就是当从公司名右边界开始向左搜索到某个位置时，无法在当前位置断定这不是公司名，还需要继续向左，如果无法发现其他公司名组成部分，才能断定这个串不是公司名。这种情况多表现为一些“公司类型名＋公司名后缀”，例如食品公司，医药公司，电脑公司，等等，它们不能单独作为公司名，但是可以和其他成分结合起来作为公司名，例如三元食品公司，联想电脑公司，等等。“公司名不完全禁止库”收录了“食品公司”之类不完全禁止词串。

四、识别策略

公司名识别策略的整体结构如下图所示：



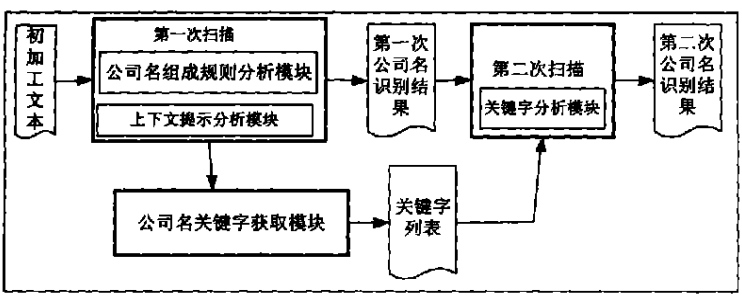
原始新闻文本首先进入分词系统。该分词系统已做了特殊的改造，添加了分词专用词典，其中包括地名词典、部分公司名词典和股票名词典。部分公司名词典中收集的是金融新闻在

最近一段时间经常出现的重要的公司名称，股票名词典则收集的是深市、沪市所有的股票名称。

经过分词、词性标注之后，我们得到了初加工文本。这个文本中已经包含有对公司名识别有用的词性信息和姓名、地名、部分公司名和股票名的信息。在识别系统的核心部分“公司名识别模块”中，我们采用两次扫描的方式，利用初加工文本中的有用信息，结合金融领域的特点，借助公司名知识库对公司名进行识别，得到公司名识别的最后结果。

其中，第一次扫描主要进行公司名全称的识别和左边界比较明显的公司名简称的识别，同时获取公司名关键字的信息。在第一次扫描中，主要利用了公司名组成规则分析法，主要用于分析公司名的全称和某些边界由地名或人名组成的公司名；同时也采用了上下文提示分析法，通过提示的信息进一步识别一些公司名。由于这样得到的左边界没有地名或人名作为左边界那么可靠，我们还需要借助禁止库对识别出的公司名进行排歧。第二次扫描则主要是利用第一轮识别中获取的公司名关键字进行识别，识别的对象是公司名的简称、左边界不怎么明显的公司名和某些右边界不明显的含有公司名关键字的公司名。

两次扫描过程的结构图如下：



五、实验结果与分析

本文使用的语料库由1336篇金融新闻文本构成，我们从中随机选出100篇新闻文本进行封闭测试，实验结果如下：

文本数目	测试点个数	识别出公司名个数	正确	错误	准确率%	召回率%	F_1
100	1179	1082	1053	29	97.3	89.3	93.1

上表中 F_1 是 $\beta=1$ 的 F 测试。

F 测试的公式为：$F=\frac{(\beta^2+1)RP}{\beta^2R+P}$。$R$ 是召回率(Recall)，P 是准确率(Precision)。

其中两次扫描后的实验结果分别为：

	识别出公司名个数	正确	错误	扫描后的准确率%	扫描后的召回率%	扫描后的F1
第一次扫描	647	631	16	97.5	53.5	69.1
第二次扫描	435	422	13	97.3	89.3	93.1

由此可以看出采用两次扫描过程进行公司名识别是非常有效的。

同时，我们还对40篇新闻文本进行了开放测试，实验结果如下：

文本数目	测试点个数	识别出公司名个数	正确	错误	准确率%	召回率%	F_1
40	597	591	371	220	62.8	62.1	62.4

我们对识别结果中的错误进行了分析,发现错误主要有以下几种类型:

1.公司类型名收集不全,导致有的简称识别不出来。

假如在公司类型名库中没有"生物工程"这个词,则碰到"四川新希望生物工程有限公司"提取的公司名关键字就是"新希望生物工程",而不是"新希望",就导致"新希望"单独作为公司名简称出现时识别不出来。随着公司类型名库的不断扩充,相信这个问题可以得到比较好的解决。

2.某些公司名既没有左边界,有没有上下文信息,导致不能正确识别。

如"绿得公司没有取得有关证照",没有可用的上下文信息,并且"得"字所标注的词性属于公司名禁止词性,这就导致"绿得公司"不能正确识别。我们试图通过局部统计的办法来解决,但由于金融新闻文本往往并不大,公司名无论全称还是简称的出现频率一般都不高,因此效果并不理想。

3.公司名简称的识别方法仍不完全。

某些公司名确实没有关键字,比如在文中先出现了"中国四川商贸股份有限公司",后面再出现"商贸股份有限公司"时是特指"中国四川商贸股份有限公司",但系统无法将该简称识别出来。这个问题目前尚未有较好的解决办法。还有一些公司名简称是由组成公司名的名词中的第一个字组成的,即前面提到的"缩写"类型,如"三明啤酒公司"简称"三啤"。这个问题将在进一步的工作中解决,可以考虑抽取公司名每个组成词的第一个字,组成公司名字串,在遇到该字串或其子串时认为它们有可能是公司名简称。

六、结束语

在金融领域的信息抽取问题中,如何正确识别文本中出现的公司名是一个非常重要的问题。本文在对金融领域中公司名的结构及其在文本中的出现环境进行了深入研究的基础上,建立了六个识别用的知识库,并提出了一个基于两次扫描过程的识别策略。经过初步实验,结果表明我们的识别策略是很有效的。

参 考 文 献

[1] 张小衡,王玲玲.中文机构名称的识别与分析.中文信息学报,1997,11(4):21—32

[2] Hsin-Hsi Chen, et. al. Description of the NTU System Used for MET2. In Proceedings of 7th Message Understanding Conference, Fairfax, VA, 19 April-1 May, 1998

[3] Erik Peterson. A Chinese Named Entity Extraction System. http://epsilon3.georgetown.edu/petersee/Chinesene.html. 1999

[4] Keh-Jiann Chen & Chao-jan Chen. Knowledge Extraction for Identification of Chinese Organization Names. In Proceedings of ACL workshop on Chinese Language Processing, 2000, 15—21

[5] N.A. Chinchor. MUC-7 Named Entity Task Definition. In Proceedings of the Seventh Message Understanding Conference. 1998

[6] N.A. Chinchor, E. Marsh. MUC-7 Information Extraction Task Definition. In Proceedings of the Seventh Message Understanding Conference. 1998