一种基于条件随机场的中文公司名识别方法

哈寅晨　孟凡坤

（北京工业大学多媒体与智能软件北京重点实验室　北京　100124）

【摘　要】随着信息化的发展，在智能信息处理领域，对自然语言处理的要求在不断提高，其中命名实体识别是一项极其重要的研究课题。本文在对信息产业新闻本文深入地研究和分析的基础上，总结出了公司名称的基本特点，分别针对公司名全称和简称，设计了不同的两种标注方式，并提出了一种基于条件随机场的双模型两次扫描识别策略，第一次扫描使用公司名全称识别模型，同时提取出公司名关键字；第二次扫描利用第一次扫描中提取出的公司名关键词改善分词和词性标注结果，在此基础上使用公司名全简称识别模型对公司名进行识别。最终的实验结果表明这种识别方法是有效的。

【关键词】命名实体识别；信息抽取；公司名；条件随机场

中图分类号：TP391.43　　文献标识码：A　　文章编号：1009-6833（2014）03-013-02

Method for Chinese Company Name Recognition Based on Conditional Random Fields

Ha Yinchen，Meng Fankun

Abstract: With the development of information society, the recognition of named entity plays a signification role in intelligent information processing.Based on the investigations and analysis of the IT news articles, the structure features and contextual constraints were obtained.In this paper, after a careful distinction of company names into two categories, i.e.full names and abbreviated names, two corresponding tagging methods are designed to represent this dichotomy and used to annotate a training corpus.This training corpus is then fed to a double-scan CRF-based company name identification system.In the first scan, full names and the keyword of the company names are recognized and extracted.In the second scan, the full names and the abbreviated names are identified based on the optimized segmentation and POS tagging result benefited from the first scan.The experimental results prove the effectiveness of this recognition method.

Keywords: Named Entity Identification; Information Extraction; Company Name; Conditional Random Fields

0　引言

命名实体识别对于很多自然语言处理领域的任务，如信息抽取，信息检索和自动文摘等而言，，是一项非常重要且基础的技术[1]。命名实体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。和第一个任务相比，后面两个子任务几乎完全可以靠几种模式匹配完成，要简单得多。因此，名字的识别（ENAMEX）是命名实体识别研究的重点。

文献[2]以人工总结的公司名构成规则和六个知识库为基础，通过两次扫描实现对文本中公司名的识别。这种方法虽然可以达到一定的准确率，但是覆盖的范围有限，仅仅依靠规则的方法很难正确覆盖自然语言中出现的所有语言现象[3]。

本文在文献[2]的基础上，提出了一种基于条件随机场（Conditional Random Fields，CRF）统计机器学习模型的公司名识别方法，在公司名的识别方面进行了有效的探索。

1　公司名特点分析和总结

公司名属于“定语+名词性中心词”型的名词短语，简称定名型短语，从宏观上看，是一种偏正复合名词，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。公司名的中心语重要集中在“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。另外，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作用的。在研究了大量的真实文本之后，我们发现在公司名中，有些词和有些词性是明显不会作为公司名的组成部分的。

表 1　公司名简称分类

简称类型	全称	简称
仅公司名关键字	深圳市一达通企业服务有限公司	一达通
	北京华胜天成科技股份有限公司	华胜天成
地名+公司名关键字	上海方正数字出版技术有限公司	上海方正
公司名关键字+公司类型	北京久其软件股份有限公司	久其软件
	海南海航航空信息系统有限公司	海航信息
公司名关键字+公司名后缀	美国苹果股份有限公司	苹果公司
地名+公司名关键字+公司名后缀	印度塔塔信息技术有限公司	印度塔塔公司

公司名的出现情况有两种：全称和简称。公司名的全简称的对应关系如表 1 所示。由此可以看出，公司名关键字的识别，对于公司名简称的识别具有非常重要的意义。

2　基于条件随机场的识别

2.1　条件随机场

条件随机场（CRF）模型最早是由 Lafferty 和 McCallum 在 2001 年提出，是一种用于在给定输入结点值时计算指定输出结点值的条件概率的无向图模型[4]。假定 O 是一个值可以被观察的“输入”随机变量集合，S 是一个值能够被模型预测的“输出”随机变量的集合，且这些输出随机变量之间通过表示依赖关系的无向边连接起来。如果用 C（S，O）表示这个图中的团的集合，CRF 将输出随机变量值的条件概率定义为与无向图中各个团的势函数（potential function）的乘积成正比：

其中，表示团 c 的势函数。当图形模型中的各输出被连接成一条线性链的特殊情形时，CRF 假设在各个输出结点之间存在一阶马尔科夫独立性，二阶或更高阶的模型可以按照类似的方法扩展。若让表示被观察的输入数据序列，让表示一个状态序列，在给定一个输入序列的情况下，线性链的 CRF 定义状态序列的条件概率为：

其中，f 是一个任意的特征函数，是每一个特征函数的权值，归一化因子为：

条件随机场模型不同于产生式模型，它可以使用丰富的、彼此重叠的观察序列的特征，而且不需要很严格的前提假设；同时，不同于最大熵马尔科夫模型等概率模型，它不是对单个标记归一化之后再进行全局搜索，而是在整个观测序列上求解一个最优的标记序列，避免了标记偏见问题。因此，条件随机场模型本身非常适合用于中文命名实体识别等这样的任务。

2.2　标注方式

针对中文公司名的识别，我们将句子的分词结果和词性信息二者作为识别公司名的重要的特征信息，用于条件随机场模型[5]。

由于公司名全称具有相对明显的左右边界词特征[6]，所以区别于由 Ramshaw 和 Marcus 提出的 BIO 标注方式，即 B（begin，开始）、I（internal，内部）和 O（other，其他）。本文针对公司名全

称的结构特点，以及应对从全称中提取公司名关键字的需要，提出了一种 BKTEO 的标注方式，即 B（begin，开始）、K（keywords，公司名关键字）、T（type，公司类型）、E（end，公司名后缀）和 O（other，其他），构成标注集合。标注示例如下表所示：

由	p	O
上海	ns	CN-B
玖	m	CN-K
峰	q	CN-K
数码	n	CN-T
科技	n	CN-T
有限公司	n	CN-E
提供	v	O

针对于公司名简称，采用 BCEO 的标注方式，即 B（begin，开始）、C（continue，延续）、E（end，结束）和 O（other，其他），构成标注集合。标注示例如下表所示：

和	cc	O
玖峰	nz	CN-B
科技	n	CN-C
有限公司	n	CN-E
总裁	n	O

2.3 特征提取

条件随机场模型可以利用丰富的、彼此重叠的特征，所以在应用中一个非常重要的问题就是如何针对特定的任务为模型选择合适的特征集合，用这样的特征集合表示复杂的语言现象[7]。相对于隐马尔科夫模型只能利用中心词的前n个词作为上下文信息的弱点，条件随机场模型能够同时使用中心词的前 n 个词和后 m 个词作为该词的上下文信息，这样，中心词的最终标记不仅与前面词语的信息相关，还与其后的词语相关，更加接近实际情况。

针对中文公司名的识别，我们设置了大小为 5 的上下文观察窗口，利用平行输入的词形 W（word）和词性 P（Part of Speech）信息，对于待标注的词，其标注结果依赖如下特征：

1,, （1）

2,, （2）

3, （3）

这三个式子分别表示，待标注词的标注结果依赖于其所在位置前后两个词的词形和自身的词形，依赖于其所在位置前后两个词的词性和自身的词性，以及其前一个词的标注结果。

3 识别策略

公司名识别策略的整体结构图如下图 1 所示：

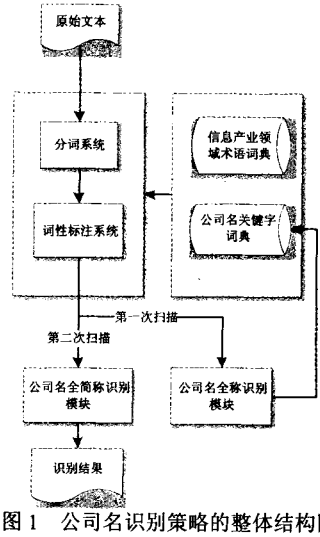

图 1 公司名识别策略的整体结构图

原始新闻本文首先进入分词和词性标注系统，该系统已经经过了初步的改造，加入了信息产业领域常用概念和术语，以改善分词和词性标注的效果。另外，加入了部分公司名关键字，用于识别公司名简称。

第一次扫描主要进行公司名全称的识别和公司名关键字的提取。在第一次扫描时，原始文本经过分词和词性标注之后，进入到“公司名全称识别模块”，被识别出的公司名全称中表示为的部分被提取出来，作为公司名关键字加入到公司名关键字词典中，并以“专有名词”（nz）作为其在字典中的词性标注，以此改善第二次扫描时的分词和词性标注结果。

第二次扫描则主要是利用第一次扫描中获得的公司名关键字信息和改善后的分词和词性标注结果，识别包含有公司名关键字的公司名简称。

4 实验结果和分析

本文使用的语料库来自互联网的信息产业新闻网站，共收集了 13283 篇。从中随机选出了 100 篇新闻文本，对公司名全称采用 BKTEO 的标注方式进行人工标注，作为训练集，用于训练识别公司名全称的条件随机场模型。另外，同样的对这 100 篇新闻文本，对所有的公司名实体（包括全称和简称），采用 BCEO 的标注方式进行人工标注，作为训练集，用于训练识别公司名全简称的条件随机场模型。

对这 100 篇新闻文本进行封闭测试，公司名全简称识别实验结果如下：

文本数目	100
测试点个数	1099
识别出公司个数	903
正确数	870
错误数	33
准确率	96.3%
召回率	82.2%
F1	88.7%

我们对结果中错的识别进行了分析，总结如下：

（1）对于公司名类型的识别，非常依赖训练集的标注数量，导致有些公司名不能识别。

（2）公司名关键字的提取的错误会传递到第二遍扫描，即造成公司名全简称识别错误。

（3）有些公司名的简称，特别是国企简称，其全称本身通常不带有关键字，如“中国电子科技集团公司”简称为“中电集团”，其中就不包含任何公司名关键字，给识别工作带来了困难。

5 结束语

本文介绍了一种基于条件随机场的公司名的识别方法。首次提出了利用 CRF 统计模型自动标注的方法提取公司名关键字。经过初步试验，结果表明我们的识别方法是可行有效的。下一步的工作是对本文所提出的方法进行改善，这包括扩充训练集的数量，对全称识别结果进行后处理，进一步过滤掉错误的识别，以提高第二遍扫描的准确率。

参考文献：

[1]孙镇，王惠临.命名实体识别研究进展综述[J].现代图书情报技术，2010，06：42-47.

[2]王宁，葛瑞芳，苑春法，等.中文金融新闻中公司名的识别[J].中文信息学报，2002，16（2）：1.

[3]廖先桃.中文命名实体识别方法研究[D].哈尔滨工业大学，2006.

[4]Lafferty，John D.；McCallum，Andrew；Pereira，Fernando C.N.：Conditional Random Fields：Probabilistic Models for Segmenting and Labeling Sequence Data.In：Proceedings of the Eighteenth International Conference on Machine Learning（ICML 2001），Morgan Kaufmann Publishers，2001，pp.282-289.

[5]张祝玉，任飞亮，朱靖波.基于条件随机场的中文命名实体识别特征比较研究[C].第四届全国信息检索与内容安全学术会议论文集.北京：出版者不详，2008：111-117.

[6]邱莎，王付艳，申浩如，段玻，阿圆，丁海燕.基于含边界词性特征的中文命名实体识别[J].计算机工程，2012，13：128-130.

[7]黄利科，刘群.基于条件随机场的中文产品名自动识别方法[J].计算机应用研究，2008，25（10）：1829-1831.

作者简介：

哈寅晨（1986—），男，北京工业大学硕士研究生，研究方向：知识工程。

孟凡坤（1986—），男，北京工业大学硕士研究生，研究方向：知识工程。